Support Vector Machine-based Fuzzy Systems for Quantitative Prediction of Peptide Binding Affinity

Volkan Uslan
BSc, MSc

A thesis submitted in partial fulfilment of the
requirements for the degree of Doctor of Philosophy
at De Montfort University

March, 2015
Leicester

I would like to dedicate this thesis to my mother, Ayse Uslan, and my father, Ismail Uslan.

Acknowledgements

The author would like to offer his sincerest gratitude to his first supervisor, Dr Huseyin Seker, who gave him the opportunity to study bioinformatics and helped, supported, and encouraged him all the way through his PhD. The author would also like to thank his second supervisor Professor Robert I. John for his invaluable support, assistance and guidance. Moreover, the author would also like to thank his family and knows without their invaluable support, this study would not have been successful.

Abstract

Reliable prediction of binding affinity of peptides is one of the most challenging but important complex modelling problems in the post-genome era due to the diversity and functionality of the peptides discovered. Generally, peptide binding prediction models are commonly used to find out whether a binding exists between a certain peptide(s) and a major histocompatibility complex (MHC) molecule(s). Recent research efforts have been focused on quantifying the binding predictions.

The objective of this thesis is to develop reliable real-value predictive models through the use of fuzzy systems. A non-linear system is proposed with the aid of support vector-based regression to improve the fuzzy system and applied to the real value prediction of degree of peptide binding. This research study introduced two novel methods to improve structure and parameter identification of fuzzy systems. First, the support-vector based regression is used to identify initial parameter values of the consequent part of type-1 and interval type-2 fuzzy systems. Second, an overlapping clustering concept is used to derive interval valued parameters of the premise part of the type-2 fuzzy system.

Publicly available peptide binding affinity data sets obtained from the literature are used in the experimental studies of this thesis. First, the proposed models are blind validated using the peptide binding affinity data sets obtained from a modelling competition. In that competition, almost an equal number of peptide sequences in the training and testing data sets (89, 76, 133 and 133 peptides for the training and 88, 76, 133 and 47 peptides for the testing) are provided to the participants. Each peptide in the data sets was represented by 643 bio-chemical descriptors assigned to each amino acid. Second, the proposed models are cross validated using mouse class I MHC alleles (H2-Db, H2-Kb and H2-Kk). H2-Db, H2-Kb, and H2-Kk consist

of 65 nona-peptides, 62 octa-peptides, and 154 octa-peptides, respectively. Compared to the previously published results in the literature, the support vector-based type-1 and support vector-based interval type-2 fuzzy models yield an improvement in the prediction accuracy. The quantitative predictive performances have been improved as much as 33.6% for the first group of data sets and 1.32% for the second group of data sets.

The proposed models not only improved the performance of the fuzzy system (which used support vector-based regression), but the support vector-based regression benefited from the fuzzy concept also. The results obtained here sets the platform for the presented models to be considered for other application domains in computational and/or systems biology. Apart from improving the prediction accuracy, this research study has also identified specific features which play a key role(s) in making reliable peptide binding affinity predictions. The amino acid features "Polarity", "Positive charge", "Hydrophobicity coefficient", and "Zimm-Bragg parameter" are considered as highly discriminating features in the peptide binding affinity data sets. This information can be valuable in the design of peptides with strong binding affinity to a MHC I molecule(s). This information may also be useful when designing drugs and vaccines.

Contents

List of Figures

List of Tables

List of Publications produced from PhD Thesis

Papers published

V. Uslan, H. Seker and R.I. John, "A Support Vector-Based Interval Type-2 Fuzzy System," IEEE International Conference on Fuzzy Systems, pp. 2396-2401, 6-11 July 2014, Beijing, China.
(**Chapter-6**)

V. Uslan and H. Seker, "Support Vector-based Fuzzy System for the Prediction of Mouse Class I MHC Peptide Binding Affinity," the 13th IEEE International Conference on BioInformatics and BioEngineering, pp. 1-4, 10-13 November 2013, Crete, Greece.
(**Chapter-5**)

V. Uslan, and H. Seker, "Support Vector-Based Takagi-Sugeno Fuzzy System for the Prediction of Binding Affinity of Peptides," the 35th IEEE International Conference on Engineering in Medicine and Biology Society, pp. 4062-4065, 3-7 July 2013, Osaka, Japan.
(**Chapter-5**)

Papers in preparation

V. Uslan and H. Seker, "Modelling Non-linear System in the Post Genome Era: Quantitative Prediction of Degree of Peptide Binding by using Support Vector based Fuzzy System".
(**Chapter-5**)

V. Uslan and H. Seker, "Survey on Quantitative Prediction in Bioinformatics, Systems and Computational Biology".
(**Chapter-2**)

V. Uslan, H. Seker and R.I. John, "Modelling Non-linear System in the Post Genome Era: Quantitative Prediction of Degree of Peptide Binding by using Support Vector based Type-2 Fuzzy System".
(**Chapter-6**)

Abbreviations

AA	**A**mino **A**cid
MHC	**M**ajor **H**istocompatibility Complex
DNA	**D**eoxyribo**n**ucleic **A**cid
RNA	**R**ibo**n**ucleic **A**cid
NGS	**N**ext **G**eneration **S**equencing
FS	**F**uzzy **S**ystem
T1-FS	**T**ype-**1** **F**uzzy **S**ystem
T2-FS	**T**ype-**2** **F**uzzy **S**ystem
IT2-FS	**I**nterval **T**ype-**2** **F**uzzy **S**ystem
TSK-FS	**T**akagi **S**ugeno **K**ang **F**uzzy **S**ystem
FCM	**F**uzzy **c**-**M**eans
MF	**M**embership **F**unction
FOU	**F**ootprint **O**f **U**ncertainty
UMF	**U**pper **M**embership **F**unction
LMF	**L**ower **M**embership **F**unction
HCM	**H**ard **c**-**M**eans
HIE	**HIE**rarchical
MCFS	**M**ulti **C**luster **F**eature **S**election
SVM	**S**upport **V**ector **M**achine
SVR	**S**upport **V**ector **R**egression

A List of Symbols

symbol	name
x	(primary) variable
X	universe of discourse
A	fuzzy set
$\tilde{A}$	type-2 fuzzy set
u	degree of membership
$\mu_A(x)$	membership function
$\mu_{\tilde{A}}(x,u)$	type-2 membership function
J_x	primary membership
f	firing level
$\prod$	product t-norm
a	the coefficients of consequent
q^2	coefficient of determination
ρ	spearman rank correlation coefficient

Chapter 1

Introduction

1.1 Motivation

The first human genome was sequenced more than a decade ago [1], [2] and has become available for further scientific research studies. It is undoubtedly a great discovery and the completed sequence contained more than three billion base pairs. One aspect of the project is that not only did the project get the benefit of advanced molecular biology methods, but also computational methods. The project relied heavily on the computational efforts, particularly during the final phase. Hence, one consequence of this great project is that the computer aided biological research is and will be essential.

The completion of sequence of human genome means a new era of research studies began which is referred to as post-genome era. Advances in the genome-technology have yielded vast amount of data during this era. An intense analysis was required in order to discover biological knowledge and derive clinical information from the underlying data. The developments in biological complex problems and genomic technologies with huge amount of data inevitably require the connection of computational methods and life sciences. Promising solutions and approaches were offered by the algorithms dedicated to solve particular problems in biological systems. Nevertheless, data produced by these technologies challenges research studies, forcing them to develop new strategies to better analyse and model the information and integrate them with biological systems [3].

The need to better analyse and retrieve valuable information in biological data sets bloom the field of bioinformatics. It is an interdisciplinary research area one step towards the

better analysis of biological data sets using computational methods and appropriate software tools in order to address the complex bio-problems. As being a young field, bioinformatics contain some uncertainty in its definition. It may mean different to different people. In the post-genome era, it is no denying that bioinformatics will take the center stage in contributing to modern biology and even become the major part of it. Janet Thornton, a professor at Cambridge University, says that "if the computational tools are well designed, then gradually all biologists will become applied bioinformaticians at some level" [4].

The bioinformatics data sets are often challenging in the post-genome era. Not only they are vast and high-dimensional, but also measured data is often incomplete and contains uncertainty. Therefore, computational methods under the development aim at reducing noise and high-dimensionality as well as dealing with the incompleteness and uncertainty in such data sets.

Prediction of binding affinity is one of the application domains in bioinformatics where data is often complex, uncertain and high-dimensional. Human reasoning can mostly process low-dimensional data sets as compared to computers that can capable of processing big amounts of data in high-dimensions. Conventional methods are often not adequate and solely limited to human reasoning capability. Moreover, information produced in wet-labs is extremely limited. Therefore, the computer aided prediction of binding affinity is crucial in order to leverage the analysis of these biological data sets. This thesis mainly addresses modelling non-linear system in the post genome era and concerned quantitative predictions related to bioinformatics and systems biology. The range of application domains in computational biology is broad as reviewed in the literature review of this thesis. From these wide range of topics, this research study focuses on the quantitative prediction of peptide binding affinity being regarded as one of the difficult modelling problems in bioinformatics.

In this research study a novel fuzzy system than can efficiently model a non-linear system is proposed. Fuzzy systems are able to model uncertain and imprecise knowledge and forms a structure for representing human reasoning. Usually, fuzzy systems can be constructed by obtaining the knowledge from human experts. Nonetheless human experts may not be available all the time, and building a model using a classical non-linear system with a limited prior knowledge is often difficult [5]. Among the various

fuzzy systems, Takagi-Sugeno-Kang (TSK) is commonly used for modeling complex systems [6], [7]. TSK fuzzy systems can be combined with other methods, particularly learning methods, and enhanced with learning and adaptation capabilities [8]. SVR concept is incorporated in our model with TSK-FS to better train the consequent part of the TSK-FS. In addition, fuzzy clustering has been used to derive the premise part of fuzzy system to approximate the membership functions that characterise each fuzzy set found in the rule-base and to identify structure of the fuzzy model [9], [10].

In the consequent section (Section 1.2) an overview of amino acids, peptides and proteins is presented. Section 1.3 introduces the peptide binding affinity problem. Contributions of the PhD study is provided in Section 1.4. Finally, the structure of thesis is explained in Section 1.5.

1.2 Amino Acids, Peptides and Proteins

An amino acid is a bio-molecule that contains an amine (NH) and a carboxylic (CO) acid group. A peptide bond joins carboxyl acid group of one amino acid to amine group of another as shown in Fig. 1.1.

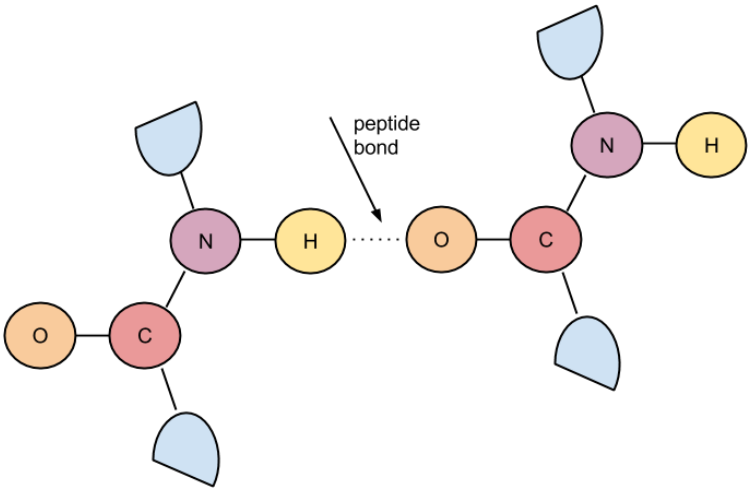


FIGURE 1.1: The formation of a peptide bond through the linking of atoms.

A peptide is a small molecule as compared to protein with a two or more amino acids attached to each other by peptide bonds. A peptide has a molecular structure similar to protein.

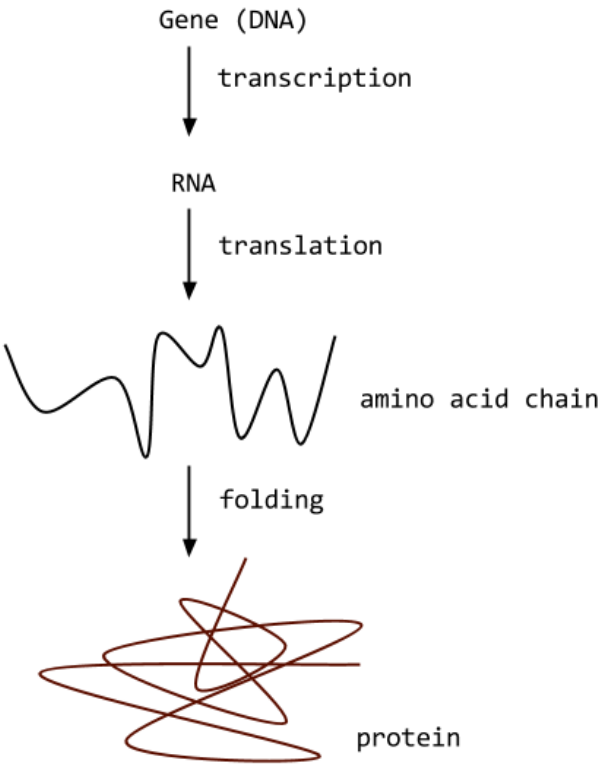


FIGURE 1.2: The course of protein production.

There are twenty different amino acids. The list of amino acids and their side chain information are given in the Table 1.1. Each amino acid contains information about its specifics such as molecular weight, volume, polarity and composition.

Proteins are made up of amino acids, attached to each other by peptide bonds. The tertiary structure and biological activities of proteins are often decided through the use of sequence of amino acids. Twenty different amino acids are bound together in a variety of combinations forming a folded structure and yielding proteins that have distinct three dimensional structure and biological functions. Fig. 1.2 depicts the course of a protein production.

1.3 Peptide Binding Affinity

Our body is always under the attack of unwanted guests or intruders, namely bacteria, fungi, parasites or viruses. Apart from these pathogens, it is also possible that healthy cells may become tumor cells [11]. Security and protection mechanisms are needed in order to fight and deal with such cases. It is gratifying that our immune system is in

charge. White blood cells (leukocytes) in the immune system protect our body from infection. T-cells, B-cells and natural killer cells are the principal types (lymphocytes) of white blood cells. Immune system recognises antigens that invades into our body and triggers a protection response [12]. The adaptiveness of the immune system allows different response mechanisms for different kind of antigens.

The main response mechanism on the cell level is the cytotoxic T-cells which are responsible to initiate response mechanism when the cell is infected by a virus or become malignant. When the infection happens whether it is cancer or viral, the proteins remained as the cause of the infection resides within the cell. Through a digestion procedure performed by proteoses these proteins converted into a number of peptides. The generated peptides are translocated to the endoplasmic reticulum of the cell. These translocated peptides are bound to MHC molecules. The 3D structure of a peptide binding to MHC

TABLE 1.1: List of amino acids with their symbolic representations and side chain information.

Amino Acid	**3-Letter**	**1-Letter**	**Side Chain Description**
Alanine	**Ala**	**A**	non-polar and neutral
Arginine	**Arg**	**R**	polar and basic
Asparagine	**Asn**	**N**	polar and neutral
Aspartic acid	**Asp**	**D**	polar and acidic
Cysteine	**Cys**	**C**	polar and neutral
Glutamine	**Gln**	**Q**	polar and neutral
Glutamic acid	**Glu**	**E**	polar and acidic
Glycine	**Gly**	**G**	non-polar and neutral
Histidine	**His**	**H**	polar and basic
Isoleucine	**Ile**	**I**	non-polar and neutral
Leucine	**Leu**	**L**	non-polar and neutral
Lysine	**Lys**	**K**	polar and basic
Methionine	**Met**	**M**	non-polar and neutral
Phenylalanine	**Phe**	**F**	non-polar and neutral
Proline	**Pro**	**P**	non-polar and neutral
Serine	**Ser**	**S**	polar and neutral
Threonine	**Thr**	**T**	polar and neutral
Tryptophan	**Trp**	**W**	polar and neutral
Tyrosine	**Tyr**	**Y**	polar and neutral
Valine	**Val**	**V**	non-polar and neutral

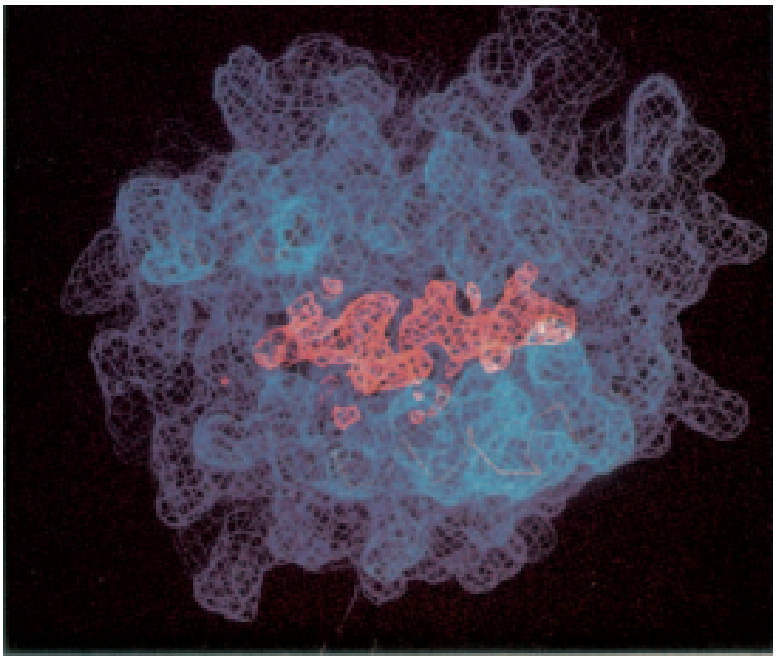


FIGURE 1.3: 3D structure of peptide binding to MHC class I.

class I molecule is shown in Fig. 1.3 (figure adapted from [13]). Then the MHC-peptide complex is translocated on the surface of the infected cells so that it can be an activation signal for a T-cell receptor present at the T-cell surface [14]. These bindings have outmost importance in that they induce cellular immune responses [15]. This process is illustrated on a diagram as shown in Fig. 1.4.

Revealing the association of peptides with the MHC molecules can be crucial for a drug design and development. A common assessment to elicit these associations is to find peptide binding affinity. One of the most challenging and complex aspect of the peptide binding is the prediction of protein-peptide binding affinity.

Peptide binding prediction models are commonly used to find out whether a binding exists between peptide and MHC molecule [16]. The prediction methods that are commonly used of this kind are BIMAS [17] and SYFPEITHI [18]. Many other prediction methods are also available such as RANKPEP [19] and SVMHC [20] which are based on the position specific scoring matrices (PSSMs) and SVM to find out whether a peptide might bind, respectively. They are often able to determine the tendency and strength of the bindings in order to save time as well as experimental efforts. The qualitative models further improved and focused on modeling to classify binders as strong and weak binders rather than determining the existence of a binding as binders or non-binders [21], [22],

[23]. Recent research efforts that are of particular interest in this application domain have been focused on quantifying the binding predictions [24], [25].

This thesis is concerned with the binding affinity problem in which high-dimensionality of data sets and uncertainties involved in them are common issues. Proposed models aim at predicting quantitative peptide binding affinities rather than peptides might bind or not such as SYFEPEITHI does. Finding a feasible solution to this bioinformatics problem remains an open issue. Moreover, there is still need for new methods, which take into account the complexity of the problem. Fuzzy systems are highly capable of dealing with the uncertainties in the measurements therefore it is considered they can be useful in dealing with such a problem as this.

1.4 Contributions of the PhD Study

Since it is believed that fuzzy systems are capable of tackling with complex problems, this thesis suggests quantitative predictive fuzzy models that can provide a feasible solution to the binding affinity problem. The research studies in this thesis that are considered to contribute to the literature are summarised as follows:

- A support vector based fuzzy system is proposed and applied to the binding affinity prediction problem which is one of the complex modelling problems in bioinformatics due to the diversity of peptides discovered. The results clearly suggest a positive impact of the fuzziness concept on SV-based methods. The improved generalisation ability of the fuzzy system is experimented and tested with two validation methods. The results are clearly better than the presented results in the literature. (conference papers are published [26], [27] and journal article is in preparation [28])

- A novel clustering approach is developed to identify premise parameter values for type-2 fuzzy systems. There is no straight-forward method in order to find the initial parameters of type-2 fuzzy membership functions. These parameters are commonly arbitrarily initialised in the generation process of rule-based type-2 fuzzy systems. Overlapping clustering framework is proposed to reveal the parameters of interval type-2 membership functions. The experiments showed that the

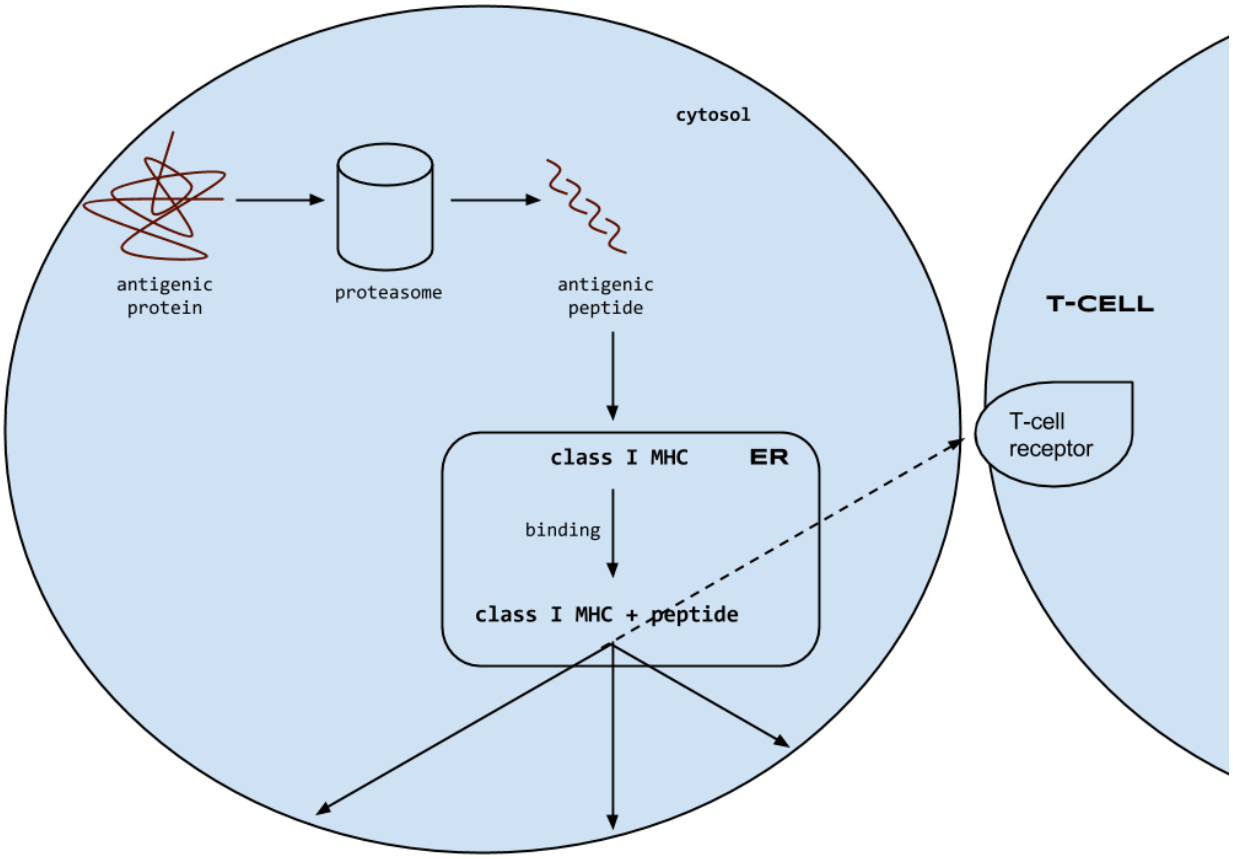


FIGURE 1.4: The process of the peptide binding.

proposed approach yielded better determination of parameters of interval type-2 fuzzy membership functions as compared to the arbitrary initialisation of these membership functions. (journal article is in preparation [29])

- A novel type-reduction and defuzzification approach is developed for the SV-based type-2 fuzzy modelling. In this approach, the support vector based regression is used to identify the structure and parameter values of the consequent part and integrated with a closed mathematical form where the type-reduction is not necessary. (conference paper is published [30] and journal article is in preparation [29])

- An extensive review that covers the quantitative prediction problems and proposed solutions to them in the fields of bioinformatics and systems biology, is conducted. Regression-based methods that are used to confront presented problems, are presented. (journal article is in preparation [31])

1.5 Thesis Structure

The rest of this thesis is organised as follows:

Chapter 2 reviews the literature that relates quantitative prediction in bioinformatics and systems biology. This literature review focuses on describing related biological background and the state-of-the-art of the field and latest developments in quantitative prediction in bioinformatics and systems biology. Comparative analysis of the developed methods is discussed to focus and address various kinds of biological complex problems. Furthermore, regression based methods that are used in the proposed models in the literature are explored. The review chapter will be turned into a review journal paper as there doesn't seem to be such a comprehensive review in this growing field.

Chapter 3 presents the background theory for the construction of SVR-based fuzzy systems. The proposed models of this thesis are composed of fields of computational intelligence such as clustering methods, fuzzy system modelling and regression-based methods and hybridisation of these that can address quantitative nature of biological complex problems.

Chapter 4 presents the construction of peptide data sets through the AA indices of which the descriptions and their scales collected from literature. The pre-processing of the bioinformatics data sets through the feature extraction and selection process are described intensively to provide insight view of the characteristics of the data sets that are dealt with.

Chapter 5 presents and characterises an SVR-based type-1 fuzzy system that encompasses a series of experiments to demonstrate the robustness of this experimental methodology on separate peptide binding affinity data sets and mouse class I alleles. The improvements in comparison with the literature for both data sets are presented.

Chapter 6 presents the development of a type-2 fuzzy system that is based on overlapping clustering concept for determining the structure of premise part. Furthermore, SVR-based regression is used for initializing the coefficients of the consequent part. A closed mathematical form for type-reduction and defuzzification is incorporated to the SVR-based type-2 fuzzy modelling. Preliminary results demonstrate the ability of SVR-based type-2 fuzzy system framework in predicting real-values of peptide bindings.

Chapter 7 discusses and concludes this research study, emphasizes strengths and weaknesses, and presents contributions and future works.

Chapter 2

Literature Review

2.1 Introduction

High-throughput technologies such as next generation sequencing technologies in life sciences generate big biological data in variety of application domains. The data generated is exponentially increasing and often high-dimensional, complex and non-linear. Computational methods are therefore needed in order to ease the organization and analysis of this kind of data and help derive clinically and biologically meaningful information.

There are three main methods commonly applied in the analysis of post-genomic data. They are clustering, classification, and quantitative prediction. Clustering methods such as (e.g. fuzzy c-Means clustering) is generally applied to unlabelled data (e.g. microarray gene expression profile analysis [32]). In order to partition data into small subsets, similarity/dissimilarity of the data samples are considered. The other method is classification (e.g. sum classifier, naive bayes classifier) to be able to develop a predictive model capable of distinguishing pre-labelled classes (e.g. cancer vs. control [33]). The third method is quantitative prediction where the output was generally continuous or discrete real values. One example to quantitative prediction in the post-genome era is the binding affinity of peptides. However, this is not only a predictive method (e.g. linear or non-linear regression) but also the attribute selection highly effects the outcome of such methods.

This chapter reviews the literature and highlight the importance of the quantitative prediction in the research studies of bioinformatics and systems biology. The keyword

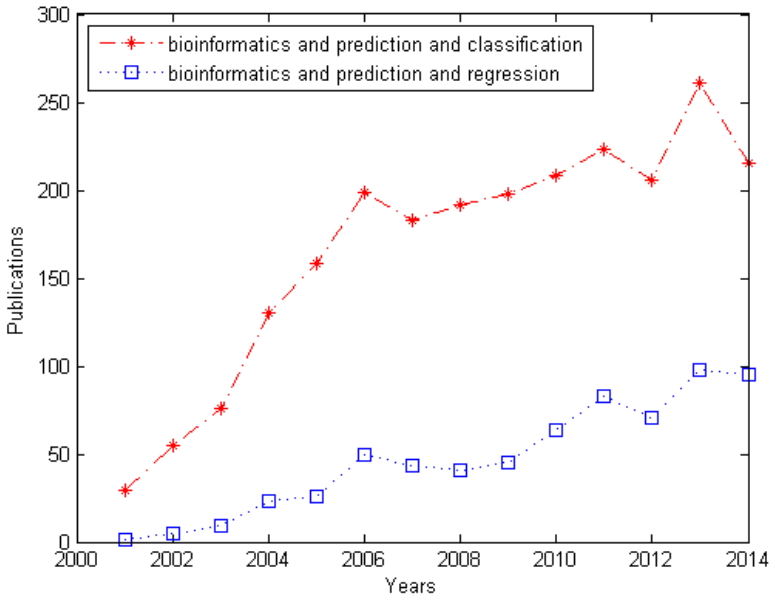


FIGURE 2.1: Number of publications per year in PubMed related to the prediction studies in bioinformatics based on classification and regression.

sets; "systems biology and regression", "bioinformatics and regression", "computational biology and prediction and regression", "systems biology and prediction and regression", "bioinformatics and prediction and regression" were used to reveal the papers from the well-known academic research databases such as Scopus, Web of Science, and PubMed. More than five hundred papers were revealed to carry out the survey but the challenge is to find out which of these studies actually were related to the quantitative prediction as most of the papers in databases were irrelevant or mainly related to the classification and clustering studies in bioinformatics.

The keywords containing classification and regression are searched separately and compared with each other. According to PubMed, the number of publications per year for the prediction studies in bioinformatics based on classification and regression is shown in Fig. 2.1. As it is clearly seen from the graph, there is a lack of quantitative prediction studies in the new era of post-genome biology as compared to classification. It should also be noted that the number of publications rose gradually from the early 2000s until present. The literature suggests that classification have been extensively studied whereas there seems a considerable smaller number of studies in the quantitative prediction. The remaining graphs of these keyword sets are presented in Appendix E.

The next section of this chapter, Section 2.2, reviews the state-of-the-art of quantitative prediction problems in bioinformatics and systems biology for various application domains. In Section 2.3, regression-based methods used in order to tackle the presented problems are briefly described. Section 2.4 provides an overview of the feature selection and reviews its use in quantitative prediction problems that have high-dimensional data sets. In Section 2.5, the importance of fuzzy systems in bioinformatics is briefly presented. Finally, Section 2.6 concludes the chapter with a final remark.

2.2 Application Domains in Bioinformatics and Systems Biology

There exists a variety of application domains in bioinformatics research studies. This section groups and reviews quantitative prediction problems into four different application domains. They are computational omics studies, systems biology, structural bioinformatics, gene expression.

2.2.1 Computational Omics Studies

Computational omics studies are the research studies in biology having the suffix -omics, which may be proteomics, genomics, metabolomics, or transcriptomics. This section presents widely used quantitative prediction research studies in computational omics studies from the selected literature (Table 2.1).

Proteomics is an emerging field concerned with the proteins expressed in an organism [34]. The studies in this omics field focus on identifying all the proteins expressed in the cells or tissues. Mass spectrometry is the method of choice widely used in order to identify and detect proteins [35]. The information in this area of research requires large-scale study and is often different from the information provided from DNA or RNA sequences [36].

A digestion procedure takes place in order to form the peptides from the proteins using enzymes such as trypsin. Mass spectrometry identifies these peptides and proteins within the biological mixture. The analysis of mass spectra involves revealing the amino acid composition of a peptide and later proteins were identified from the peptide groups.

The protein inference problems come from that these peptides can not directly related with their correct proteins due to the fact the existence of degenerate peptides and one-hit wonders. Protein inference problem can be formulated as a Logistic regression task that predicts the probability of the identified peptides with their belonging proteins [37]. ProteinLasso is a method based on peptide detectability and used as a constrained Lasso regression problem to formulate the protein inference problem [38]. Peak intensity prediction gets the use of regression methods including SVR and Linear regression and peak intensities in the measured mass spectrometry are predicted in order to identify proteins by comparing them from a database of known proteins [39], [40], [41]. Shah et al proposed a model having a set of amino acid descriptors to predict ion mobility drift times for the identification of peptides using two regression approaches, Partial Least Squares (PLS) and SVR [42].

Mass spectrometry cannot reveal all the proteins that may exist in a sample but only a portion of them [43]. The accuracy and interpretability of mass spectra is crucial in order to identify proteins. One approach that helps to improve the understanding of spectrometry data is the prediction of spectrum peak intensities using the existing molecular descriptors [44].

One of the important features of a protein is its melting temperature as it can be used particularly in efforts for drug design and development. Goronia et al collected the melting temperature of 230 proteins varying between 25°C and 113°C. They used Neural Networks (NN) and Neuro-Fuzzy methods separately to predict melting temperature of a protein from its amino acid composition [45].

Intrinsic disorders in proteins or protein regions aid understanding fundamental processes occurring in protein folding and function. Yan et al used SVR to predict intrinsic disorder on proteomic scale based on the protein sequence [46].

Genomics is the study of groups of genes in large-scale. There has been an exponential growth of data collected for genome wide association studies during last decade. Bioinformatics is heavily used in order to derive meaningful information from these genome-wide data sets. A single-nucleotide polymorphism (SNP) is the variation of a single position within the DNA sequence among individuals in a population. When a SNP occurs in a gene, it may lead a different composition of its corresponding amino acid sequence, leading to more than one allele. Although many of SNPs may not lead to

a disorder, but some of them are closely related with particular diseases. Imputations of single-nucleotide polymorphisms can be predicted using regression models. Huang et al used v-SVR to estimate the quality score of imputations of SNPs with unknown true genotypes [47].

Eukaryotic cells have wrapped sections of DNA which are called nucleosomes. Revealing nucleosome organisation is important as it provides insight information about transcription regulation. One of the factors that affect nucleosome positions is the DNA sequence. Zhang et al estimated linear factors with Linear regression and non-linear factors with SVR to predict nucleosome occupancy statistically based on di-nucleotide features of the DNA sequence [48]. Rube et al proposed a model of statistical positioning that uses Linear regression to calculate variance structure of nucleosome locations in individual genes [49].

MicroRNAs (miRNAs) are the small fragments of RNA (approximately 21 bp in length). An miRNA can interact with its corresponding target messenger RNA (mRNA) and inhibits the translation of mRNA into a protein due to imperfect binding between them. Muniategui et al uses Lasso regression for predicting miRNA-mRNA interactions [50]. Small interfering RNA (siRNA) with a length between 21 and 25 bp binds to its target mRNA causing the mRNA to degrade and cleave. This process is important and research studies focus on inhibiting or silencing gene expression in order to find prospective therapeutic solutions for cancer disease in particular. Liu et al used Ridge regression for the prediction of siRNA efficacy prediction [51]. Jiang et al used Random Forest regression to quantitatively estimate siRNAs efficiacy values [52].

During the transcription process, transcribing DNA into RNA, gene expression is regulated mostly by some specific proteins, namely transcription factors. These proteins have DNA-binding domains that help them to interact with some distinct DNA fragments called enhancer or promoter sequences. Mordelet et al used regression based model for the transcription factor-DNA binding specifity [53]. Their model contained features based on the occurrences of higher-order k-mers at various positions within or near the transcription factor binding sites.

Copy number indicates the number of copies of a given gene or parts of sequence in the whole genome [62]. Alterations in DNA copy number may indicate progress in severe disease such as cancer. These alterations are often caused from the genetic events in the

case of extreme variations in contiguous parts of the genome. Therefore, revealing DNA copy number alterations is crucial in order to follow the progression of human cancers in specific [63]. The whole genome partitioned into segments in order to find out and quantify copy number variations exists between contiguous segments. Many regression-based models including Lasso and Quantile regression proposed to analyse DNA copy number data and derive alterations that exist in such data [57], [58].

Compos et al used Lasso based model to quantitatively predict genetic values for complex traits [55]. Chen et al used Linear regression to predict causative genes for the discovery of diseases [56]. Cosgun et al uses a mixture of regression methods including SVR, Random Forest, and Regression Tree in order to predict necessary warfarin dose requirements in a cohort of African Americans [54].

Studies on protein-ligand complex and its scoring function gives valuable information regarding drug discovery. Ballester et al proposed a scoring function using Random

TABLE 2.1: Selection of widely used quantitative prediction research studies in computational omics.

Ref.	**Method**	**Application Domain**
[47]	Support vector regression	imputed genotypes
[48]	Linear regression/SVR	nucleosome occupancy
[38]	Lasso	protein inference
[44]	Artificial neural networks	protein inference
[50]	Lasso	miRNA-mRNA interactions
[51]	Ridge regression	siRNA efficacy analysis
[40]	Support vector regression	protein inference
[41]	Support vector regression	protein inference
[39]	Linear regression	protein inference
[37]	Logistic regression	protein inference
[42]	Mixture of regression methods	sequence analysis
[54]	Mixture of regression methods	genetics and population analysis
[53]	Support vector regression	transcription factor DNA binding affinity
[45]	Neural networks	melting temperature of a protein
[55]	Bayesian regression	quantitative traits
[49]	Linear regression	nucleosome occupancy
[56]	Linear regression	gene inference
[57]	Lasso	copy number alterations
[58]	Quantile regression	copy number alterations
[46]	Support vector regression	intrinsic disorder
[59]	Random Forest regression	molecular docking
[60]	Partial least squares	protein - ligand binding affinities
[61]	Support vector regression	cancer cell sensitivity

Forest to implicitly acquire binding effects of protein-ligand complexes to analyse the outcomes of the molecular docking [59]. Deng et al used PLS in order to predict protein-ligand binding affinities [60].

In modern oncology, prediction of a response of a cancer disease to a therapy may provide crucial insight information that may lead to the design of a personalized medicine. Menden et al proposed a computational framework using Random Forest and Neural Network separately based on genomic and chemical properties to predicting cancer cell sensitivity to drugs [61]. The study not only suggests identification of new drug design opportunities but also it is useful for personalized medicine associating genomic traits of patients to drug sensitivity.

2.2.2 Systems Biology

Systems biology is the field of study concerned with the understanding of interactions and predicting dynamical behaviour of biological components such as molecules and cells. Computational models are proposed and quantitative measurements are used in order to ease the tediousness of understanding the complex and dynamic behaviour of interacting biological components of living systems. Thus, systems problems of biology could be better studied, leading to proper design of drugs that can effectively bind to its biological target. This section presents widely used quantitative prediction research studies in systems biology from the selected literature (Table 2.2).

Gene regulatory networks (GRN) inferring is a reverse-engineering process in bioinformatics in order to unravel gene regulation system in a cell. Many microarray experiments produce different gene expression data. On the contrary, the genes that will be discovered are much less than the experiments being conducted. One common consequence is that the model set up is high-dimensional and can suffer from over-parameterization. There are some reviews on inferring gene regulatory networks that provide challenges in this area as well as overview common modelling schemes and applied computational methods [64].

Chan et al proposed a Least Angle regression (LARS) based model for GRN inference on a time-series microarray data of *Schizosaccharomyces pombe* yeast-cell cycle genes and the model produced biologically relevant GRN and important insight information related

to yeast cell-cycle regulation [65]. Regulatory networks found are biologically relevant and functionally correct. Xiong et al decomposed the GRN inference problem among genes and for each target gene, the expression level is predicted using Linear regression from the expression level of a potential regulation gene [66]. Xun et al inferred molecular interactions in biological systems using a Bayesian model averaging for Linear regression [67]. Andrec et al estimates the connection coefficients from noisy perturbation responses using Total Least Squares and show that the accuracy of the network structure depends not only on the noise level but on the strength of the interactions within the network [68]. Bayar et al formulates reverse-engineering genetic networks as a Multiple Linear regression (MLR) problem [69]. Qin et al uses an extended version of Lasso to infer gene regulatory network in mouse embryonic stem cells [70]. Wang et al reconstructs gene network using Lasso which uses prior information [71]. Supper et al predicts the expression level of a gene using Multiple Linear regression from a minimal combination of genes which are considered as probable regulators for that gene when unraveling GRN [72]. Yeung et al identifies a network which is sparse using Robust regression from a family of candidate networks constructed by singular value decomposition [73]. Brouard used Output Kernel regression to derive a protein-protein interaction network [74]. Qabaja used Lasso-based method to reveal functional interactions between miRNAs and diseases using miRNA gene signature [75]. Berthoumleux et al proposed a Linear regression approach in order to infer metabolic network models [76]. Castellini used a Linear regression method to reveal biological network regulations from time series [77].

Strength of binding affinity between biomolecule interactions is important for understanding biological processes happening in our body. There can be many types of biomolecular interactions. Protein-peptide interactions are one type of such interactions essential to initiate necessary responses to protect the host during his lifetime. Peptides bind to MHC proteins over the course of cell activities. Although there are potentially large numbers of peptides, they are often limited in size due to the difficulty of identification of bindings to MHC molecules. Therefore, a recent bioinformatics problem, peptide binding affinity prediction gets the aid of computational methods to ease the identification process of those peptides and to what degree that bindings can occur.

Liu et al proposed a quantitative modelling method based on SVR, namely SVRMHC, for an accurate prediction of mouse class I peptide-MHC binding affinity [25]. Subsequently, SVRMHC is used to construct and validate prediction models for over 40

MHC alleles [78]. Doytchinova and Flower studied on human MHC allele HLA-A*0201 and proposed a model to predict continuous binding affinities using Multiple Linear regression [79]. Giguere et al proposed a peptide-protein binding affinity predictor based on Ridge regression with a reasonable accurate binding affinity prediction of any peptide to any protein [80]. Demir et al used L1/L2 regularization to predict regression based typical biological problems provided from Comparative Evaluation of Prediction Algorithms contest [81]. Ivanciuc and Braun used several regression based methods in order to predict peptide-MHC binding affinities and compare them to each other [82]. Hattotuwagama et al proposed an iterative self-consistent Partial Least Squares based additive method in order to predict class II MHC-peptide binding affinity [83]. Guo et al proposed a novel string kernel and uses SVR to predict class II MHC-peptide binding affinity [84]. Shao et al used SVR to predict PDZ domain-peptide interaction from primary sequence [85]. Doytchinova et al used Linear regression to fit actual binding affinities of test peptides to the predicted ones [86]. In a further work, Doytchinova et al used MLR in order to assess their additive method for the prediction of binding affinity [87]. Guan et al proposed a method called MHCpred and used PLS to evaluate its statistics [88]. Subsequently, MHCpred is enhanced with the addition of mouse class I models and the removal of computational constraints and become MHCpred 2.0 [89]. Previously, the prediction server contained human class I and II models. Bordner et al proposed methods called RTA [90] and MultiRTA [91] and used L1/L2 regularization to select a subset of initial parameters in order to avoid overfitting from their model. Chang et al uses PLS to predict class II MHC-peptide binding based on peptide length [92]. El-manzalawy used Multiple Instance regression to predicting MHC-II binding affinity [93].

Determining the protein-protein interaction affinity is a significant research area of systems biology where binding affinity takes place in order to infer real status of the protein-protein interaction networks. However, not many promising solutions suggested to address the problems of protein-protein interactions including binding affinity and structure of those interactions. Proteins interact each other and with other biological molecules to perform high level biological tasks. A protein to protein interaction (PPI) network, also known as protein interactome, is a graph that is formed by a set of vertices corresponds to proteins and a set of edges correspond to physical interactions between the pairs of proteins. Protein interaction networks may provide valuable observations

about the modularity of cellular processes and the interpretation of protein functions [97]. Over the last decade more protein interactions data become available as a result of research on finding complete genome sequences particularly on model living organisms including *Escherichia coli*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae* [98]. Thus analysing protein interactions help more to fully understand the cell mechanism [99], [100]. Furthermore they help understand the modularity of cell activities and how proteins regulate and support each other in a protein interaction network. Recent reviews describe the advances in computational methods

TABLE 2.2: Selection of widely used quantitative prediction research studies in systems biology.

Ref.	**Method**	**Application Domain**
[66]	Linear regression	GRN inference
[69]	Multiple linear regression	GRN inference
[68]	Linear regression	GRN inference
[65]	Least angle regression	GRN inference
[74]	Ridge regression	PPI inference
[71]	Lasso	GRN inference
[70]	Lasso	GRN inference
[75]	Lasso	miRNA-disease asssociation
[76]	Linear regression	metabolic network modelling
[72]	Linear regression	GRN inference
[73]	Robust regression	GRN inference
[67]	Linear regression	molecular interactions
[94]	Support vector regression	protein-protein binding affinity
[95]	Support vector regression	protein-protein binding affinity
[96]	Statistical potentials	protein-protein binding affinity
[25]	Support vector regression	protein-peptide binding affinity
[80]	Ridge regression	protein-peptide binding affinity
[81]	L1/L2	protein-peptide binding affinity
[83]	Partial least squares	protein-peptide binding affinity
[84]	Support vector regression	protein-peptide binding affinity
[85]	Support vector regression	protein-peptide binding affinity
[87]	Linear regression	protein-peptide binding affinity
[88]	Additive method	protein-peptide binding affinity
[86]	Partial least squares	protein-peptide binding affinity
[90]	Lasso	protein-peptide binding affinity
[91]	L1/L2	protein-peptide binding affinity
[92]	Partial least squares	protein-peptide binding affinity
[78]	Support vector regression	protein-peptide binding affinity
[82]	Mixture of regression methods	protein-peptide binding affinity
[93]	Multiple instance regression	protein-peptide binding affinity
[89]	Additive method	protein-peptide binding affinity

for the analysis of biological networks in the post-genomic era which infer functional modules and functional annotation of proteins [101], [102].

Li et al proposed an SVR based method in their studies that takes into account binding contributions implicitly as it is difficult to express them in the practice of modelling protein-protein binding affinity [94], [95]. Su et al studied structure-derived statistical potentials aiming at prediction of binding energy of protein-protein interactions [96].

2.2.3 Structural Bioinformatics

One of the fields of bioinformatics widely studied is the structural bioinformatics or computational structural biology. This section presents widely used quantitative prediction research studies in structural bioinformatics from the selected literature (Table 2.3). One main branch of structural bioinformatics is to analyse and predict biomolecular structures, in particular protein structures. Proteins are essential building blocks of a cell and the biological processes are mediated and regulated through proteins and interactions of proteins. Protein structure prediction is a challenging problem in bioinformatics that helps elucidating the structure of 3D and function of a protein.

Predicting the 3D structures solely from amino acid sequences is a difficult task. The first step achieving this purpose is to reveal the secondary structure of the protein or the solvent accessibilities of protein‘s structure [103], [104], [105], [106], [107] . This way can be more convenient as it provides simpler 1D projections of the secondary structure to work on to reveal complicated 3D structure [108]. Solvent accessibility is one of the important attributes of amino acid residues that aids predicting structures of proteins. Surface area of a macromolecule which is accessible to a solvent is referred to as solvent-accessible surface area or in short accessible surface area (ASA). ASA is generally measured in square angstroms which is a standard metric in molecular biology. The prediction of solvent accessibility helps to elucidate relation between structure of a protein and its interactions [109]. The prediction finds the degree to which residues in the structure interact with the solvent molecules. Conventionally, residues can be considered as two (exposed/buried) or three (exposed/intermediate/buried) classes for the given protein structure. This burial degree of a residue helps to understand sequence-structure-function relationship and predict structural and functional properties of proteins. Nevertheless, the real value prediction is getting important due to the ill-defined classes of solvent accessibility in

real structures of proteins. The burial core residues are of crucial importance during the folding process of the protein [110]. On the contrary exposed residues help understanding proteins function as the active sites that make bound with bio-molecules found are on the surface of a protein [111]. SVR is a common regression approach to predict real values of solvent accessibility surface area in square angstroms from their amino acid sequences/primary structures [112], [113], [114], [115]. The other regression approaches for the real value prediction, also possible in this regard, include Neural Network-based regression [116], [117] and Multiple Linear regression [118], [119]. A linear dependency exists between the contribution of individual residue to folding stability of a protein and its buried solvent-accessible surface area [120]. Xu et al gets the benefit of this linear dependence and used Quadratic programming and a statistical energy function to predict solvent accessibility by performing constrained optimisation of protein stability upon burial of amino acid residues [121].

Different from the solvent accessible surface area prediction, that studies residues which are mostly on the surface of a protein, the protein burying depth prediction, as a structural descriptor, provides how residues are arranged within the inner structure of a protein and how deep they bury themselves in the formation of protein folding process. Thus, more accurate information as in the form of real residue depth values would be obtained as compared to solvent accessibility related to residues arrangement from protein sequences rather than knowing solely whether they are exposed or buried [122], [123], [124]. Accurately predicting residue depth values have many uses including folding process and recognition and functional site prediction. Protein folding determines the three-dimensional shape of a protein from its primary structure. Therefore understanding the folding mechanism of a protein will provide a valuable insight about its structure. Huang et al used Quadratic regression to predict folding rate change of a protein based on amino acid substitutions [125].

The sequence driven prediction of 3D structure of a protein and its function are a crucial task in bioinformatics due to the big difference between the number of protein structures and the number of protein sequences revealed from conducted laboratory experiments. Protein-folding problems start mostly with the secondary structure prediction from the available protein sequence. The torsion angles (Φ) and (Ψ) are commonly used to determine the backbone structure of a protein. These angles rotate around the peptide bonds. Predicting or knowing the torsion angles helps to identify the structure of a

protein due to the plane nature of linked rigid peptide bonds. The backbone angles constantly vary due to the continuous movement of proteins. Prediction is performed through the information provided from the amino acid residues. Neural Network-based regression is mainly used for improving the torsion angle prediction [126]. One of the approaches that improves the torsion angle also gets benefit from the angle periodicity [127]. Song et al uses a two-level SVR approach for an accurate prediction using the descriptors derived from the amino acid sequences [128].

As many protein structure models suggested in the literature fail to produce desired results, there is a need for experimental validation of those structures and assessment of their qualities. Many scoring functions attempt to sort and rank separate models that are driven with the same sequence. For particular application domains, however, assessment of quality of structure is crucial in order to apply the model to specific problems. There are attempts reported based on the regression for the assessment of quality of protein structures. SVR is commonly used to develop a scoring function to assess the accuracy of protein structures [129], [130]. Tondel used Multivariate regression for the prediction of homology model quality directly from the sequence alignment [131]. Yang et al developed regression equations including Linear and Logistic regressions to assess the quality of structure models of whole *Escherichia coli* proteome [132].

Seeking and finding the correct positions of residue contacts or coordination number in proteins partly characterizes protein tertiary fold structure. Each residue center has a spherical cutoff that involves residues falling inside this sphere. Determining an accurate functional relationship between amino acid sequence and the number of stabilizing contacts is crucial in predicting protein structure. Therefore, predicting the number of contacts for each residue, or coordination number is another key attribute toward predicting particularly secondary structure of a protein [133], [134]. Finding the correct positions of residue contacts in proteins help in the prediction process. As a regression task, SVR is the method of choice in general to predict this kind of prediction problem [135], [136].

Disulfide bonds are one of the structural elements within a protein that contribute the stability of the protein structure and give insight information about the proteins folding process. SVR commonly used to predict disulfide connectivity patterns in order to improve the prediction of protein secondary structure [137], [138]. Lund et al used a

Neural Network prediction approach in order to find interatomic distances in proteins [139].

The research studies are mostly focused on the prediction of protein structures. On the contrary, prediction of genomic structures are also studied such as the RNA secondary structure prediction [140]. However, protein structure predictions take the centre stage in structural bioinformatics.

TABLE 2.3: Selection of widely used quantitative prediction research studies in structural bioinformatics.

Ref.	**Method**	**Application Domain**
[105]	Support vector regression	protein secondary structure
[123]	Support vector regression	residue depth
[140]	Support vector regression	RNA secondary structure
[127]	Neural networks	backbone torsion angle
[133]	Neural networks	residue contacts
[134]	Neural networks	residue contacts
[139]	Neural networks	interatomic distance
[135]	Support vector regression	residue contacts
[121]	Quadratic programming	solvent accessibility
[116]	Neural networks	solvent accessibility
[114]	Support vector regression	solvent accessibility
[118]	Linear regression	solvent accessibility
[128]	Support vector regression	backbone torsion angle
[132]	Mixture of regression methods	quality assessment
[136]	Support vector regression	quality assessment
[131]	Multivariate regression	quality assessment
[137]	Support vector regression	disulfide connectivity
[115]	Support vector regression	solvent accessibility
[125]	Quadratic regression	folding change rate
[112]	Support vector regression	solvent accessibility
[129]	Support vector regression	quality assessment
[106]	Logistic regression	protein secondary structure
[130]	Support vector regression	quality assessment
[117]	Neural networks	solvent accessibility
[126]	Neural networks	backbone torsion angle
[113]	Support vector regression	solvent accessibility
[122]	Support vector regression	residue depth
[107]	Logistic regression	protein secondary structure
[138]	Support vector regression	disulfide connectivity
[119]	Linear regression	solvent accessibility
[124]	Support vector regression	residue depth
[104]	Linear regression	protein secondary structure

2.2.4 Gene Expression Analysis

Gene expression analysis studies and analyses a set of genes to understand the transcriptional behaviour of cell functions. It is widely used in order to subclassify the diseases, identify the key genes, and elucidate the biological pathways [141]. This section presents widely used quantitative prediction research studies in gene expression analysis from the selected literature (Table 2.4).

Microarray experiments produce gene expression profiles that contain the expression levels of thousands of genes. Cell activities in an organism can be observed by using these profiles. When there is a substantial change occurs between the profiles of an organism, this may be a sign of disease. In their proposed work, Raghava and Han studied an SVR based method to correlate and predict gene expression level from amino acid composition of a protein [142].

These gene expression data sets are huge in size and inevitably contain missing values due to the fact that resolution may be insufficient or image may be corrupted. Wang et al uses SVR as an impute method to predict the missing values that reside within the one row of certain microarray gene expression profile [143].

The microarray technology can also be used to reveal phenotypes of patients quantitatively from their gene expression profiles as well as disease studies. Fitting quantitative phenotypes becoming important in bioinformatics as it is often hard to classify samples into proper classes where high variability of individuals exists. Quantitative phenotype

Table 2.4: Selection of widely used quantitative prediction research studies in gene expression analysis.

Ref.	**Method**	**Application Domain**
[144]	Support vector regression	gene expression analysis
[145]	Support vector regression	quantitative phenotypes
[146]	Gaussian process regression	gene expression analysis
[147]	Logistic regression	molecular pathway identification
[143]	Support vector regression	gene expression analysis
[148]	Least angle regression	cancer studies
[149]	Logistic regression	cancer studies
[142]	Support vector regression	gene expression analysis
[150]	Logistic regression	cancer studies
[151]	Support vector regression	cancer studies
[152]	Support vector regression	expression noise

prediction from genotype or gene expression data can be required particularly when studying the complex common diseases in order to classify samples into their correct classes. Gui et al proposed a study related to the survival of patients that suffers from cancer after they took the chemotherapy. This study uses Least Angle regression to identify genes during the course of survival of the patient [148]. Levin et al used a Logistic regression based approach in order to identify chromosomal regions that have significant changes in gene expression in human tumors [149]. Chen et al proposed a new regularized least squares SVR for gene selection and used many data sets related to cancer [151]. Bielza et al proposed a Logistic regression method without a penalty term and applied this method to several microarray data sets for the purpose of cancer classification [150].

Guzetta et al used SVR to fit quantitative phenotypes from genotypes and used L1/L2 regularization to output the optimal weight vector [145]. Gene and pathway selection is also a challenging task in bioinformatics, in particular when they are indicative of some sort of disease. Zhang et al identifies molecular pathway with subtypes of disease using Logistic regression from gene expression profiles [147].

Some other regression based methods proposed as well in the literature that related to issues with microarray data. Liu et al estimate replicate time shifts caused by the biological development time of each replicate using Gaussian process regression from time-course gene expression data sets [146]. Myasnikova et al used SVR to address the estimation of the embryo age of a Drosophila melanogaster according to its gene expression pattern [144]. Dong et al proposed a predictive model to predict expression noise of a gene using SVR [152].

2.3 Regression-based Methods

There are many regression methods reported and applied to the various problems in bioinformatics and systems biology. In this section, commonly used ones in separate application domains, are going to be explained (Table 2.5).

Linear regression is one of the fundamental and extensively used regression methods in statistics. It uses the least squares method as an objective function to minimise the sum of residuals which is squared difference between the dependent and independent

real-values of the given data set. The method seeks to capture the relationship between multiple predictor variables and the response variable. The input and output variables are mostly denoted with capital X and Y, respectively. When only one dependent variable is used then this is a simple Linear regression. However, the models are mostly constructed in real-world problems with multiple descriptors. This is called the Multiple Linear regression. It should be stated here that the both cases involve only one response variable Y. The case of multiple response linear regression is called the Multivariate Linear regression.

Quantile regression is a regression method proposed as an alternative to commonly used Linear regression that estimates a conditional mean [153]. The method aims at estimating conditional percentile functions rather than a conditional mean. The main advantage of Quantile regression is that it is more robust against outliers as compared to the Least squares estimation.

Random Forests are a cohort of decision trees from randomly generated repeated samples of a training data set [154]. As a computational method, the Random Forests can represent information related to conditional relations between variables and can be used not only for classification tasks, but also for regression tasks as well [155]. Its regression ability is reported to yield a high-prediction accuracy as compared to its counterparts for omics data.

Least angle regression is a recent computationally efficient model selection algorithm derived from the traditional forward selection methods and different from them as it is less greedy but more useful [156]. LARS has three main properties. The first is it can implement Lasso and calculate all possible Lasso estimates for a given problem in a much faster way. The second is it can implement Forward Stagewise Linear regression and provides similar results as compared the Lasso and Stagewise. The third is it can provide a simple approximation for the degrees of freedom of a LARS estimate.

The Support Vector Machines, initially formulated by Vapnik based on statistical learning theory [157] aiming at structural risk minimisation, can be used for both continuous (Support Vector Regression) or discrete (Support Vector Classification) estimation problems. In comparison with Linear regression, SVR ensures high generalizability and performance as it is capable of tolerating errors up to a value from the expected response variables.

TABLE 2.5: Selection of widely used regression-based methods in bioinformatics.

Method	**Description**	**Advantages**	**Disadvantages**
Multiple linear regression (MLR)	captures the relationship between multiple predictor variables and the response variable	simple, widely used	difficult to model real-world problems
Quantile regression (QL)	aims at estimating conditional percentile functions rather than a conditional mean	more robust against outliers	computationally inefficient and have additional parameters
Random Forests (RF)	a cohort of decision trees from randomly generated repeated samples	effective when some of the data is missing	may suffer from overfitting
Least angle regression (LARS)	derived from the traditional forward selection methods	computationally efficient in high feature space	sensitive to the outliers
Support vector regression (SVR)	based on statistical learning theory aiming at structural risk minimisation	ensures high generalizability and performance	parameter and kernel determination

2.4 Feature Selection for Quantitative Prediction Models

Feature selection methods and application domains will be discussed in the following subsections to highlight the importance of feature selection in the study of bioinformatics and systems biology.

Feature selection aims to find the least number of dimensions (features) that contribute most to the performance and accuracy of a model. It is frequently used for data preprocessing. Feature selection helps simplify a model and alleviates the effect of the curse of dimensionality problem. It also helps better generalization and interpretation of the model. Guyon and Elisseff [158], in their methodological paper, have focused on two categories of feature selection methods, namely feature ranking methods and variable selection methods. This research study focused also on these two categories of feature selection as their wide use in the application domains of bioinformatics and systems biology. In feature ranking methods, the features are ranked by a metric. These methods apply a ranking criterion to distinguish between the variables. Those who have a good predictive power in the prediction performance of the model are ranked as top features. On the other hand, subset selection methods search for an optimal subset of features that contribute most to the accuracy. One disadvantage of feature selection methods is that an additional computational cost is involved in the preprocessing stage of the model building process. A subset of features needed to be searched and ranked in the feature space to get rid of irrelevant features. Therefore, feature selection methods are more applicable when the data set is high-dimensional and the model suffers from the effect of curse-of-dimensionality. Nevertheless, interestingly, feature selection methods themselves can be sensitive to curse of dimensionality [159]. Many of them can be prone to overfitting. There are studies related to improve the feature selection process, particularly to reduce the curse-of-dimensionality effect [160]. Therefore, one main advantage of a feature selection method amongst others is its ability to avoid from overfitting and its resistance against the effect of curse-of-dimensionality.

It should be noted here that the main concern of this thesis is to propose a predictive modelling approach for the studied bioinformatics problem. This section is added as the feature selection is the preprocessing stage of the computational predictive models that involve high-dimensionality. Rather than the supervised feature selection methods

commonly appear in high-dimensional bioinformatics applications, an unsupervised feature selection is used throughout this thesis. Compared to supervised feature selection methods, unsupervised feature selection methods do not require the target variables in the selection process. Therefore, they are less likely dependent on the target variables and more data samples - even their target variables are absent - can be used in searching for relevant information.

2.4.1 Application Domains

Scherbart et al [44] used a Neural Network approach for mass spectrometry prediction by peptide prototyping. In the proposed work, a feature selection is applied heuristically and the feature space is formed of 18 features.

In the work of Chen et al [114], a sequence based prediction of relevant solvent accessibility is presented and included a custom-selected subset of features based on Pearson correlation coefficient.

Zhang et al [122], proposed a method that predicts sequence-based residue depth using evolutionary information and predicted secondary structure. High-dimensionality of the feature set is addressed using a correlation-based feature selection.

In Compos et al [55], dense molecular markers and pedigree in the regression model to predict quantitative traits is presented. The model uses Bayesian regression coupled with Lasso to fit marker affects in the regression model from a large number of markers.

In the work of Liu et al [51], a multi-task learning method for cross-platform siRNA efficacy prediction is presented. L1-norm regularization (Lasso) is used to control the features learned in the multi-task learning process.

In the proposed work of Guzetta et al [145], the model fits quantitative phenotypes from genotypes and used L1/L2 regularization to output the optimal weight vector.

Demir-Kavuk et al [81] used a two-step regularization procedure to predict typical peptide problems provided from an online prediction contest. They used Lasso regularization for the feature selection stage of their model building process and subsequently followed Ridge regularization for the prediction stage with the use of these selected features.

In the work of Mordelet et al [53], stability selection for regression-based models of transcription factor-DNA binding specificity is presented. The features are based on the occurrences of k-mers at different positions in transcription factor binding sites. As the generated feature set from k-mers is formed of thousands of parameters and leads to overfitting in the training data, Lasso regression is used as a feature selection method.

Uslan and Seker [26], [27] proposed a support vector-based fuzzy system to predict binding affinity of peptides for various peptide data sets and mouse class I MHC alleles. To reduce the dimensionality of the large feature set that is about 5500 features, an unsupervised feature selection approach is used.

Chen et al [56] proposed a work that integrates different human omics data sources to prioritize candidate genes whose genetic bases are completely unknown. Lasso is used as to filter the irrelevant data sources by zeroing the weight of them. The remaining data sources are considered to as good data sources and used in computing candidate gene scores.

In Dong et al [152], variability in gene expression that can be used in predicting stochastic noise level is presented. This work uses the feature selection based on several criteria of mutual information [161] to select the most relevant features in predicting noise level.

2.4.2 Methods for Feature Selection in Biological Domains

In the previous section the applications in different application domains in bioinformatics and systems biology are overviewed. This section focuses on the feature selection methods used in these applications. As can be clearly seen from Table 2.6, the feature selection methods mostly used are the Lasso and correlation-based methods. This section describes them briefly.

The L1 penalty of Lasso regression eliminates irrelevant features and helps to decrease the size of the feature set [162]. The model output is often presented as a linear function of inputs. The regression aims for estimating the coefficient vector based on the least square error and the coefficient weight absolute values. At the end of the regression process, many of the absolute value of coefficient weights becomes zero. The features

TABLE 2.6: Selection of widely used feature selection methods in bioinformatics and systems biology.

Ref.	**Method**	**Application Domain**
[44]	Heuristic	Proteomics
[114]	Correlation-based	Systems Biology
[122]	Correlation-based	Structural Bioinformatics
[55]	Lasso	Genomics
[51]	Lasso	Systems Biology
[145]	Lasso	Genomics
[81]	Lasso	Systems Biology
[53]	Lasso	Genomics
[26]	Unsupervised	Systems Biology
[27]	Unsupervised	Systems Biology
[56]	Lasso	Genomics
[152]	Correlation-based	Gene Expression

having zero coefficients are eliminated as they do not have any effect on the output value of the regression process. The objective function of the Lasso regression given as follows:

$$\min \frac{\lambda}{2}\|w\|_1 + \sum_{i=1}^{n}(y_i - w^T x_i)^2 \tag{2.1}$$

where lambda is the regularization parameter denotes the trade-off between fit and sparse of inputs and w denotes the vector of regression coefficients. Based on the penalty term, as the lambda value increases the L1 norm of weight vector becomes sparser. On the other hand; as the lambda value approaches to zero, it becomes more like ordinary least squares. In the end, the solution involves zeroing out some elements of w so that a reduced feature set is obtained. Therefore, effective setting the value of the lambda parameter is important. One disadvantage of Lasso regression is that the perturbations within the training data set can negatively affect the feature set to be produced.

The correlation based feature selection is based on the linear correlation coefficient r [163], [164]. This approach filters the redundancy within the feature set yielding a subset of features. The linear correlation coefficient r, for the x and y variables, is given

as follows:

$$r = \frac{\sum_{i=1}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}(x_i - \overline{x_i})^2}\sqrt{\sum_{i=1}(y_i - \overline{y_i})^2}} \tag{2.2}$$

where x_i and y_i denote the mean values of x and y, respectively. The value of r is in the interval between -1 and +1. The strong correlation between x and y variables indicated by the higher absolute values of r. A full correlation means, the value of r is -1 or +1. A zero correlation means, the value of r is 0 indicating x and y are completely independent from each other. In the correlation based feature selection, each feature can be ranked based on the r value between the feature value and the actual output value.

2.5 Fuzzy Systems in Bioinformatics

Bioinformatics and medicine research studies generate large data sets. These data sets often involve biologically meaningful information. They are also uncertain and imprecise to some extent due to their characteristics of being complex, high-dimensional and non-linear. Computational methods are therefore required to handle and analyse this kind of data. One such method is the fuzzy systems (extensively studied in the next chapter), a computational tool capable of handling and minimizing the levels of uncertainties and imprecision. Fuzzy logic is utilized in many application domains of bioinformatics [165], [166].

2.6 Final Remark

In this chapter, the state-of-the-art of the quantitative prediction in the research studies of bioinformatics and systems biology are reviewed. As one can see in the review, variety of regression methods are used and applied in this manner. Regression methods that are commonly used in various application domains are briefly explained. The availability of the quantitative predictive solutions or those proposed as a tool that are covered in this review are presented in Table 2.7. The high-dimensionality is another concern when building the models. Feature selection methods are highly utilised in

order to eliminate such concerns. It has been noticed that in the context of regression-based models, Lasso and correlation are the feature selection methods that commonly used. It should be noted that, accuracy is much important than the computational efficiency in bioinformatics research studies. However, computational efficiency may become important in order to conduct the data analysis in the case of limited computer hardware availability.

The literature suggests that the number of research studies in the quantitative prediction are less than those studies in classification. However, an increasing trend in the number of quantitative prediction studies is observed during the course of period (from 2001 onwards). Since it is believed that many quantitative bioinformatics problems remain an open issue, more research efforts need to be directed towards such problems.

The literature review showed that support vector regression is the method of choice in various application domains of bioinformatics. To our best knowledge, there are no methods suggested in bioinformatics literature benefiting from the collective strengths of fuzzy logic and SVR. Therefore, this research study considers the cooperation of fuzzy systems with the support vector based systems in order to provide generalizability as well as minimizing the levels of uncertainties in predicting the affinities of peptide bindings.

TABLE 2.7: The availability of the reviewed quantitative predictive models in application domains of bioinformatics and systems biology.

Ref.	Method	Application Domain	Availability/Tool
[38]	Lasso	protein inference	http://sourceforge.net/projects/proteinlasso
[50]	Lasso	miRNA-mRNA interactions	http://talasso.cnb.csic.es/
[51]	Ridge regression	siRNA efficacy analysis	http://lifecenter.sgst.cn/RNAi/
[42]	Mixture of regression methods	sequence analysis	http://omics.pnl.gov/software/imPredict.php
[53]	Support vector regression	TF DNA binding affinity	http://genome.duke.edu/labs/gordan/ISMB2013
[55]	Bayesian regression	quantitative traits	http://www.genetics.org/content/suppl/2009/03/16/genetics.109.101501.DC1
[56]	Linear regression	gene inference	http://bioinfo.au.tsinghua.edu.cn/bridge
[57]	Lasso	copy number alterations	http://bioinformatics.med.yale.edu/DNACopyNumber
[46]	Support vector regression	intrinsic disorder	http://biomine.ece.ualberta.ca/RAPID
[66]	Linear regression	GRN inference	http://www.the-dream-project.org
[71]	Lasso	GRN inference	http://nba.uth.tmc.edu/homepage/liu/pLasso
[73]	Robust regression	GRN inference	https://sites.google.com/site/bmalr4netinfer/
[88]	Additive method	protein-peptide binding affinity	http://www.jenner.ac.uk/MHCPred
[92]	Partial least squares	protein-peptide binding affinity	http://malthus.micro.med.umich.edu/Bioinformatics
[78]	Support vector regression	protein-peptide binding affinity	http://svrmhc.umn.edu/SVRMHCdb
[93]	Multiple instance regression	protein-peptide binding affinity	http://ailab.cs.iastate.edu/mhcmir
[96]	Statistical potentials	protein-protein binding affinity	bioinfo.tsinghua.edu.cn/ suyu/ppepred
[89]	Additive method	protein-peptide binding affinity	http://www.jenner.ac.uk/MHCPred
[140]	Support vector regression	RNA secondary structure	http://www.tbi.univie.ac.at/∼ivo/RNA/
[127]	Neural Networks	backbone torsion angle	http://sparks.informatics.iupui.edu
[133]	Neural Networks	residue contacts	http://promoter.ics.uci.edu/brnn-pred/
[128]	Support vector regression	backbone torsion angle	http://sunflower.kuicr.kyoto-u.ac.jp/∼sjn/tangle
[112]	Support vector regression	solvent accessibility	http://birc.ntu.edu.sg/∼pas0186457/asa.html

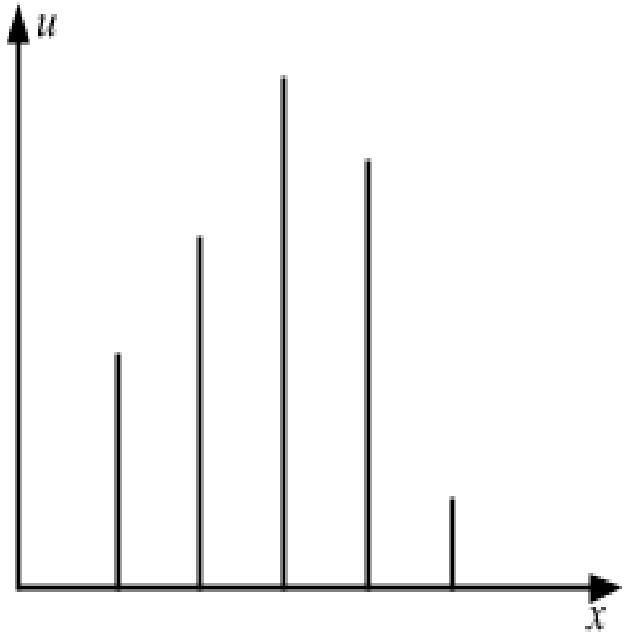
Chapter 3

Background Theory

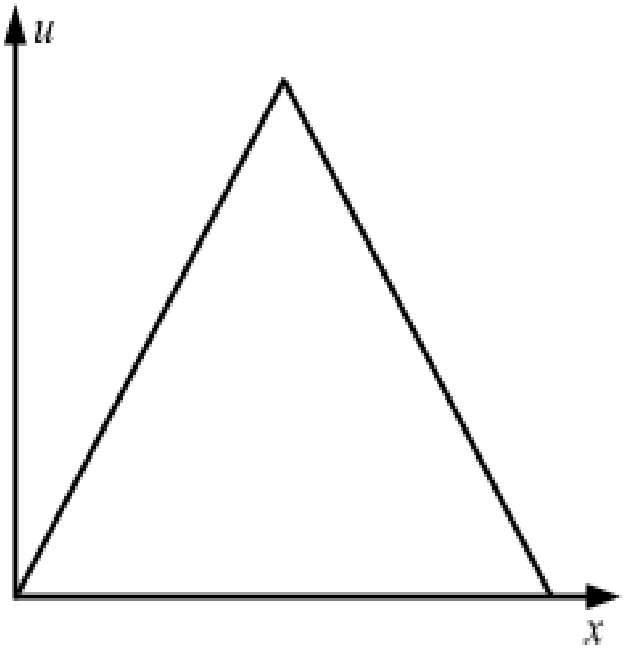
3.1 Introduction

Bioinformatics and systems biology data sets often involve uncertainty. There is no guarantee for a bioinformatician that the data set received is fully reliable. The raw data produced could be unreliable and erroneous in some degree even though thorough quality control steps were applied [167]. Furthermore, quality steps performed may not be adequate to this data set prior to initiating the bioinformatics analysis. In that sense, fuzzy systems can provide mechanisms to handle and minimise such uncertainty/unreliability for a better judgement and increase statistical power on the data sets that is dealt with.

Firstly, in Section 3.2, fuzzy logic systems are presented. Sections 3.3 and 3.4 are concerned with the structure and parameter identification of the fuzzy modelling. SVM-based regression, presented in Section 3.3, and cluster analysis, presented in Section 3.4, provide background information about the presented methods. Feature selection that is used to decrease the dimensionality of feature space is discussed in Section 3.5. Finally, Section 3.6 provides measurements used to assess the performance of the predictive models.



(a) A discrete type-1 fuzzy set.



(b) A continuous type-1 fuzzy set.

FIGURE 3.1: A type-1 fuzzy set.

3.2 Fuzzy Logic Systems

Zadeh's work in 1965 introduced a new dimension on the thinking upon the classical (crisp) set theory [168]. This new dimension is the uncertainty. Classical binary (two-valued) logic considers membership of the objects to a set in that an object could be a member or not a member of the set. Fuzzy sets bring a new dimension by relaxing the sharp boundary that exists between membership or non-membership. Therefore, it is important to understand the relationship between crisp and fuzzy sets.

The uncertainty is common in real-life, it is very hard for humans to consider everything in the sense that it is crisp (true or false). There are always thoughts beyond the two-valued logic especially when the interpretations are based on information that is incomplete, imprecise, unreliable or vague [169]. Humans express thoughts in their natural language with the use of linguistic words. Due to this fact, Zadeh introduces the linguistic variable [170] as a computing term in contrast the numerical variables which the computing is based on.

3.2.1 Type-1 Fuzzy Logic Systems

Uncertainties are often handled with a rule-based fuzzy system, namely a type-1 fuzzy system, based on a set of fuzzy sets. Fuzzy sets extends the concept of the sharp boundary that exists between membership or non-membership of elements in classical sets by enabling them to have membership degrees in the interval of 0 and 1.

A type-1 fuzzy system contains four main components: fuzzification, rule-base, inference engine and defuzzification. In the fuzzification stage type-1 fuzzy sets are generated. A type-1 fuzzy set can be either discrete or continuous. The former has discrete values of x_i where each x_i associated with a membership grade u_i (3.1) and the latter has continuous values of x and its associated membership grade u (3.2). A type-1 fuzzy set is illustrated in Fig. 3.1.

$$A = \frac{u_1}{x_1} + \frac{u_2}{x_2} + ... + \frac{u_n}{x_n} = \sum_{i=1}^{n} \frac{u_i}{x_i} \tag{3.1}$$

$$A = \int_x \frac{u}{x} \tag{3.2}$$

Among different fuzzy systems, there are two models widely used in the literature, namely Mamdani fuzzy systems [171] and Takagi-Sugeno-Kang fuzzy systems [6], [7]. The former can be designed as linguistic models and the latter as approximate models. Both models are formed of a set of if-then rules with the identical antecedent structures. However, consequent structures of these models are different.

Mamdani fuzzy systems are first designed as a set of linguistic rules obtained from human knowledge to control a steam engine and boiler combination. Antecedent and consequent structures of a Mamdani fuzzy rule is a fuzzy set. A typical Mamdani fuzzy system is illustrated in Fig. 3.2. To keep it simple, this fuzzy model has formed of two fuzzy rules where each rule comprised of two inputs (x_1 and x_2) and a single output (y). A Mamdani fuzzy rule can be defined as an IF-THEN proposition and can have the form of

$$\text{IF } \mathrm{x_1} \text{ is } \mathrm{A_1} \text{ and } \mathrm{x_2} \text{ is } \mathrm{A_2} \text{ THEN y is B} \tag{3.3}$$

where A_1, A_2 and B are the fuzzy sets. In the Mamdani model, each rule generates a consequent fuzzy set and then the final output fuzzy set is obtained by aggregating all these fuzzy sets using an aggregation method (e.g. max). The final output is obtained from the aggregate output fuzzy set. This process called defuzzification, where a fuzzy quantity is converted into a precise quantity. Several defuzzification methods suggested in the literature such as the center of maximum, the mean of maximum, and the center of area in order to resolve a single scalar quantity from the aggregate output fuzzy set [172]. One commonly used method to defuzzify fuzzy output function is the center of area (also called center of gravity) method [173], [174]. Although the center of area is computationally inefficient as compared to other two methods, it is the most applied method [175]. In this method, centroid of the aggregate output fuzzy set is calculated to find out a single output value.

The TSK fuzzy models have a linear function in the consequent part, which makes them different from Mamdani fuzzy models in which the consequent part is constructed using

membership functions. TSK fuzzy systems have been shown to form computationally more efficient model as they can work well with linear methods [176], [177], [178]. Moreover, optimization and adaptive methods are also more applicable to TSK fuzzy systems as both linear and non-linear optimisation techniques can be used to train such a system, which generally makes its construction faster. However, the design and training of the consequent part of the TSK fuzzy system is still open problem due to inefficient linear least square estimations. In addition, number of parameters to be trained for TSK Fuzzy systems is less than those in the Mamdani fuzzy systems. This increases the complexity of the Mamdani fuzzy system exponentially as number of input variables get higher, which is the case in most of biological system modelling problems.

The bioinformatics problem concerned in this thesis is related to the quantitative prediction of peptide binding affinity aiming at finding approximate numeric values of peptide bindings. As the regression analysis are widely used for predicting the binding degree of new peptides, it is considered to focus on designing fuzzy systems as TSK fuzzy systems. In addition, as relatively higher number of input variables is required to predict the peptide binding affinity, TSK fuzzy system is considered in order to avoid increasing computational complexity of the predictive model. Figure 3.3 shows a typical TSK fuzzy model with two fuzzy rules, two inputs (x_1 and x_2) and a single output (y). The rules are defined as conditional statements and can have the form of

$$\text{IF } \mathrm{x_1} \text{ is } \mathrm{A_1} \text{ and } \mathrm{x_2} \text{ is } \mathrm{A_2} \text{ THEN } \mathrm{y} \;=\; \mathrm{f(x_1, x_2)} \tag{3.4}$$

where A_1, A_2 are the fuzzy sets and $y = f(x_1, x_2)$ is a linear function in the consequent part. This function can be defined as

$$\mathrm{f(x_1, x_2)} \;=\; \mathrm{a_0} \;+\; \mathrm{a_1 x_1} \;+\; \mathrm{a_2 x_2} \tag{3.5}$$

where a_0, a_1, a_2 are the coefficients of input parameters (x_1 and x_2). In the TSK model each rule generates a crisp output and then the final output is obtained by aggregating all rule outputs. This process is called defuzzification, and the weighted average defuzzification value Y is computed as follows:

$$Y = \sum_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{f_i y_i} / \sum_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{f_i} \tag{3.6}$$

where f is the firing level of the fuzzy rule and its value is determined by using a conjunction operator, namely t-norm operator, which would usually be minimum or product, involved in the inference.

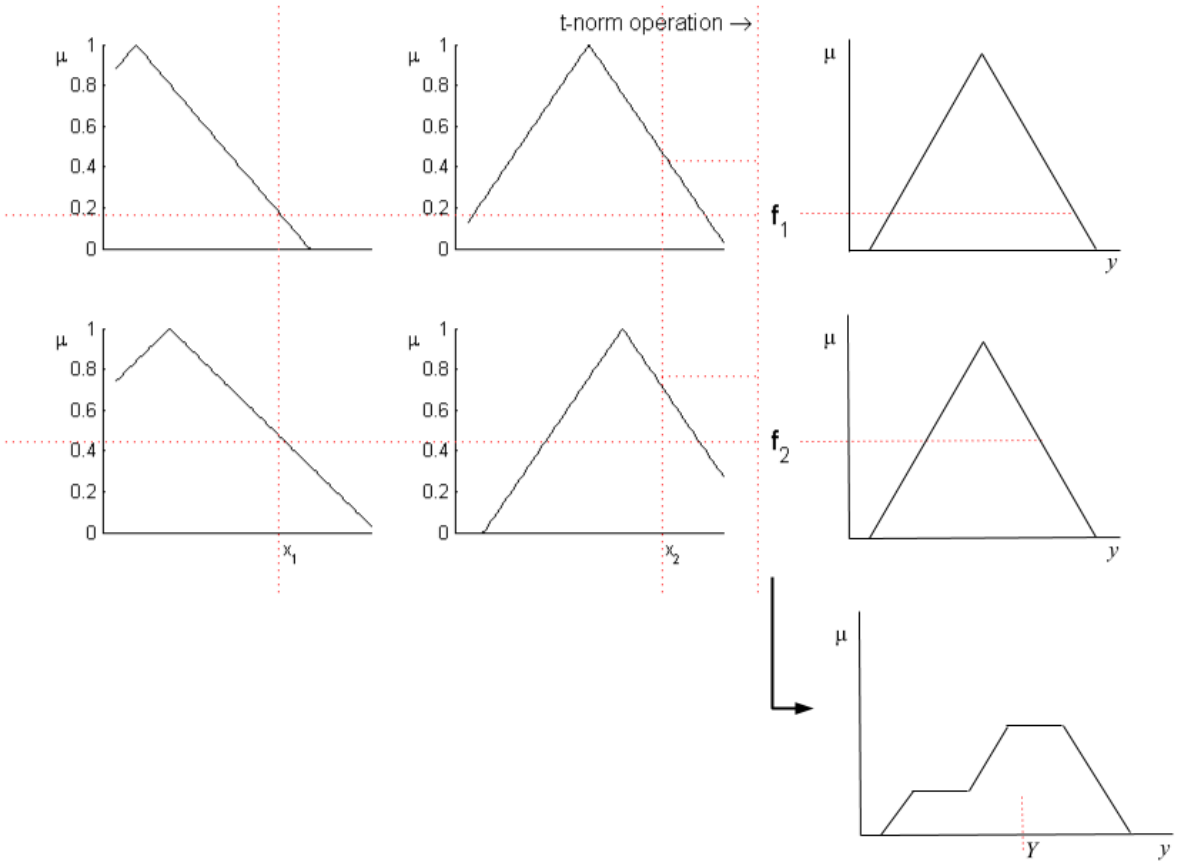


FIGURE 3.2: Mamdani fuzzy model with two inputs and single-output.

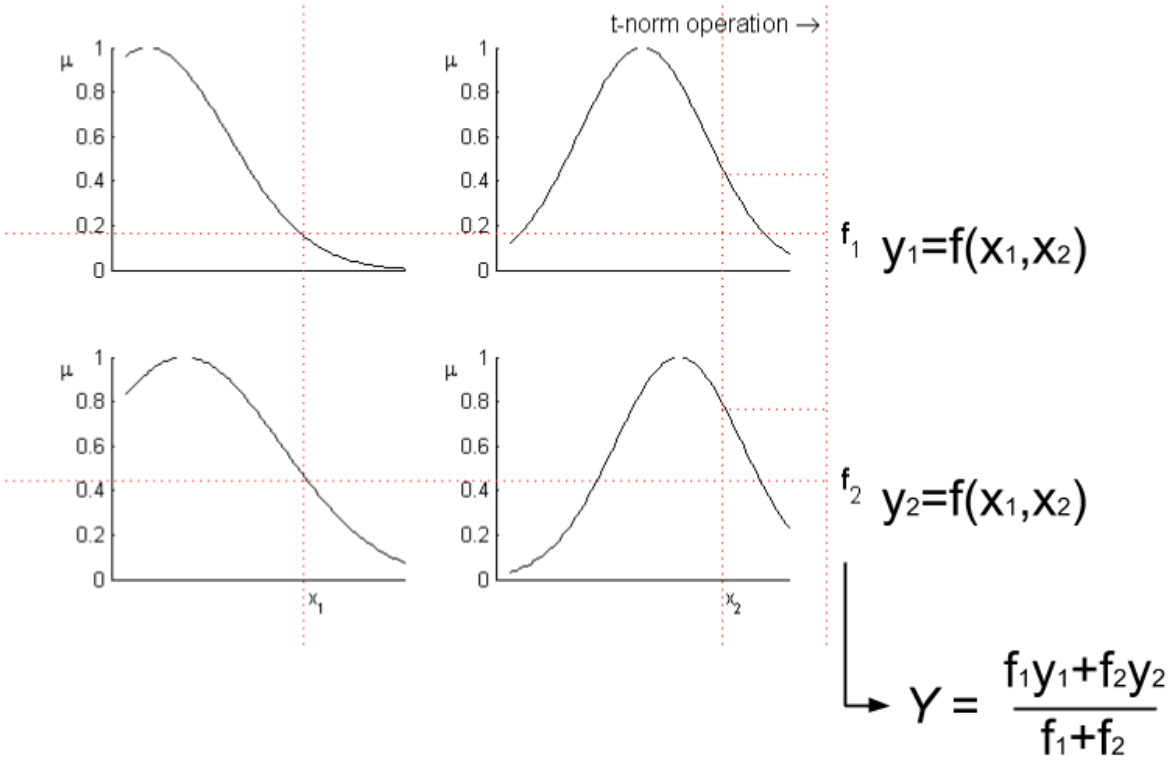


FIGURE 3.3: TSK fuzzy model with two inputs and single-output.

3.2.2 Type-2 Fuzzy Logic Systems

Type-2 fuzzy sets, which were introduced by Zadeh [179], have been shown to help better model a non-linear system and minimize the effects of uncertainties in rule-based fuzzy logic systems [180], [181], [182], [183], [184], [185]. Type-2 fuzzy sets are an extension of type-1 fuzzy sets. The membership functions that characterize type-2 fuzzy sets are themselves fuzzy. Mendel and John further improved the theoretical background of type-2 fuzzy sets and proposed a term-set to define them more precisely [186], [187], [188]. A typical type-2 fuzzy logic system structure can be shown in Figure 3.4. The definition of a type-2 fuzzy set (adopted from [186]) can be given as: A type-2 fuzzy set, denoted $\tilde{A}$, is characterized by a type-2 membership function $\mu_{\tilde{A}}(x,u)$, where $x \in$ X and $u \in J_x \subseteq [0,1]$, i.e.,

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u) \mid \forall x \in X, \forall u \in J_x \subseteq [0,1]\} \tag{3.7}$$

in which $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$. For the continuous universe of discourse, the type-2 fuzzy set can be expressed as:

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x,u)/(x,u), \;\; J_x \subseteq [0,1] \tag{3.8}$$

and for the discrete universe of discourse, the type-2 fuzzy set can be expressed as:

$$\tilde{A} = \sum_{x \in X} \sum_{u \in J_x} \mu_{\tilde{A}}(x,u)/(x,u), \;\; J_x \subseteq [0,1] \tag{3.9}$$

where $\int \int$ and $\sum \sum$ denote union over all admissible x and u, respectively.

Interval type-2 fuzzy logic systems are practical and widely used as the computations associated with the interval type-2 fuzzy sets are manageable when compared with the computational complexity of general type-2 fuzzy sets (general T2-FS) [187], [189]. Three-dimensional representations of general type-2 fuzzy set and interval type-2 fuzzy set are depicted in Fig. 3.5 and Fig. 3.6, respectively.

When the type-2 membership function, (i.e., secondary membership function) is an interval set then type-2 fuzzy logic system becomes an interval type-2 fuzzy logic system [190]. All the secondary grades $\mu_{\tilde{A}}(x,u)$ equal to 1 for an IT2-FS. IT2-FS can still be expressed as a special case of the general T2-FS. For the continuous universe of discourse,

the interval type-2 fuzzy set can be expressed as:

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} 1/(x,u), \ J_x \subseteq [0,1] \tag{3.10}$$

and for the discrete universe of discourse, the interval type-2 fuzzy set can be expressed as:

$$\tilde{A} = \sum_{x \in X} \sum_{u \in J_x} 1/(x,u), \ J_x \subseteq [0,1] \tag{3.11}$$

Figure 3.7 shows a typical representation of an interval type-2 fuzzy set. The bounded region is the footprint of uncertainty (FOU), which represents the blurring of a type-1 membership function. The FOU defines the uncertainty of an IT2-FS as:

$$\text{FOU}(\tilde{A}) = \bigcup_{x \in X} J_x \tag{3.12}$$

where $\bigcup$ denotes the union of all primary memberships. Two type-1 fuzzy sets that bound FOU are the lower and upper membership functions. The lower membership function is associated with the lower bound of FOU and the upper membership function is associated with the upper bound of FOU.

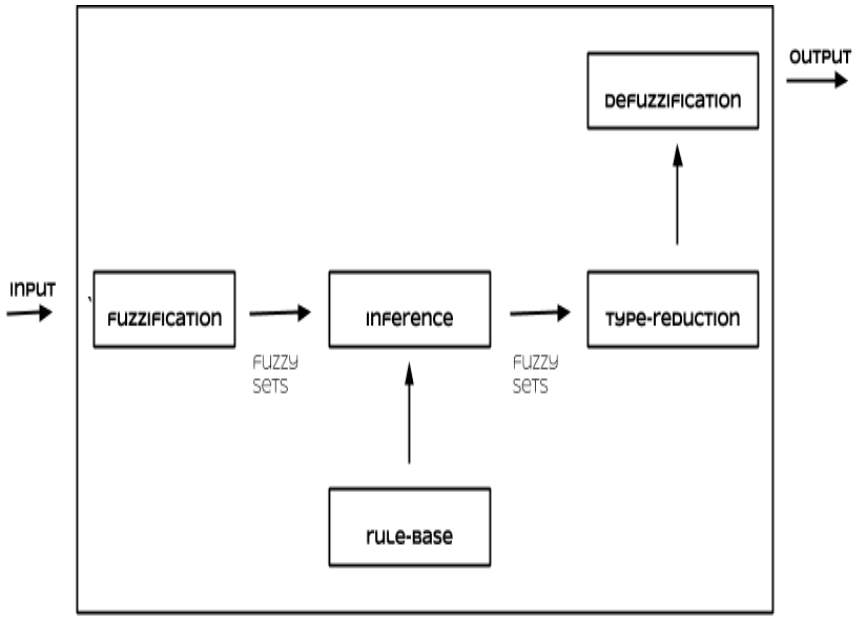


FIGURE 3.4: Type-2 Fuzzy Logic System.

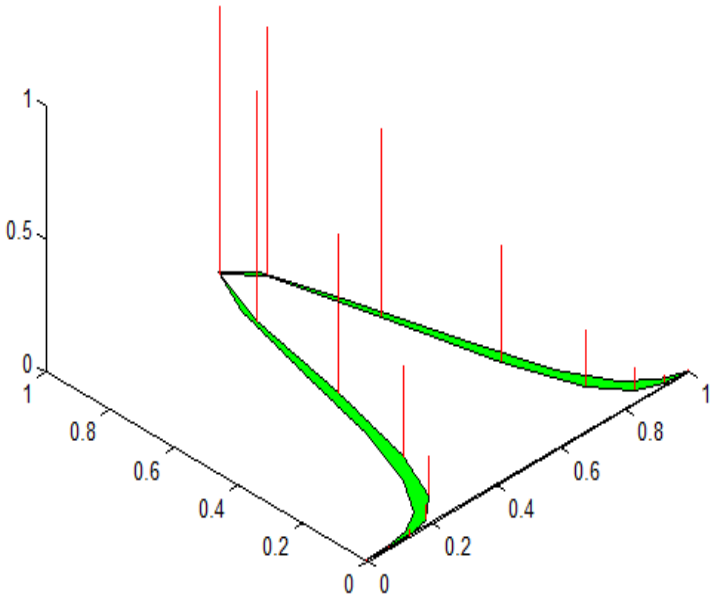


FIGURE 3.5: Example of a general type-2 membership function.

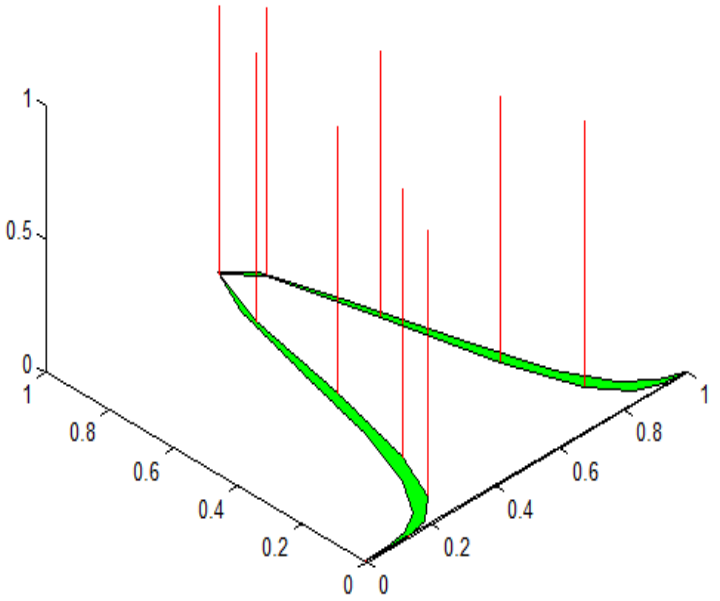


FIGURE 3.6: Example of an interval type-2 membership function.

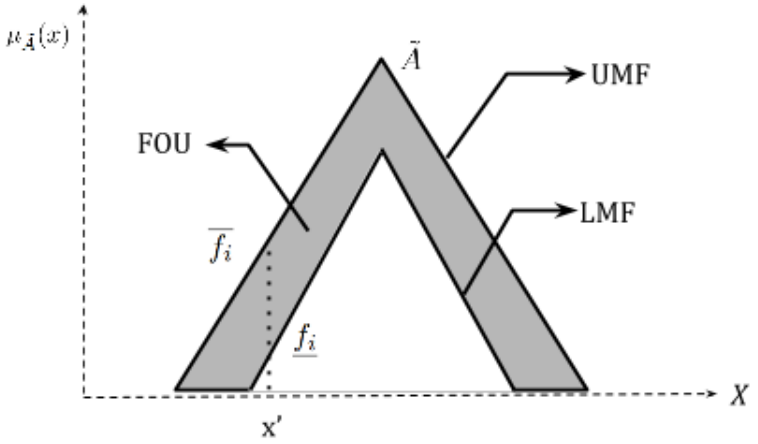


FIGURE 3.7: Interval Type-2 Fuzzy Set. UMF: upper membership function; LMF: lower membership function. The bounded region is called a footprint of uncertainty.

The output is an interval type-1 fuzzy set and represented by only left (y_l) and right (y_r) end points:

$$y = [y_l, y_r] \tag{3.13}$$

$$y = \int_{f^1 \in [\underline{f^1}, \overline{f^1}]} \cdots \int_{f^r \in [\underline{f^r}, \overline{f^r}]} 1 \Big/ \frac{\sum_{i=1}^{r} f_i y_i}{\sum_{i=1}^{r} f_i} \tag{3.14}$$

and the overall output can be calculated as:

$$Y = \frac{\mathrm{y}_l + \mathrm{y}_r}{2} \tag{3.15}$$

The TSK fuzzy model as discussed previously can be extended to its interval type-2 counterpart [191]. In this case, interval-valued fuzzy sets are used for antecedents, and a crisp output is used for the consequent part of the fuzzy rule. The fuzzy rule with two

inputs (x_1 and x_2) and single output (y) for a TSK fuzzy model has the form of

$$\text{IF } \mathrm{x_1} \text{ is } \tilde{\mathrm{A}}_1 \text{ and } \mathrm{x_2} \text{ is } \tilde{\mathrm{A}}_2 \text{ THEN y } = \mathrm{f(x_1, x_2)} \tag{3.16}$$

where $\tilde{A}$ denotes an interval type-2 fuzzy set and $y = f(x_1, x_2)$ is a linear function in the consequent part and can be defined as

$$\mathrm{f(x_1, x_2)} \ = \ \mathrm{a_0} \ + \ \mathrm{a_1 x_1} \ + \ \mathrm{a_2 x_2}$$

where a_0, a_1, a_2 are the coefficients of input parameters (x_1 and x_2). The interval type-2 membership functions $\mu_{\tilde{A}}(x)$ are used for the antecedent part of the fuzzy rule as follows:

$$\mu_{\tilde{A}}(x) = [\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)] \tag{3.17}$$

The firing strengths are determined by using the implication operator. This operator is commonly chosen as minimum or product t-norms in the inference engine. The firing strengths, computed using the product t-norm, can be in the form of

$$\overline{f} = \overline{\mu}_{\tilde{A}}(x_1) * \overline{\mu}_{\tilde{A}}(x_2) \tag{3.18}$$

$$\underline{f} = \underline{\mu}_{\tilde{A}}(x_1) * \underline{\mu}_{\tilde{A}}(x_2) \tag{3.19}$$

The defuzzified output can be computed by the Karnik-Mendel algorithms [192], [193], [194] with the steps involved from (3.13) to (3.15).

3.2.3 The Structure and Parameter Identification of a Fuzzy Model

This section presents the structure and parameter identification for two types of fuzzy models. In the next subsection, the methods used in order to identify parameters of a type-1 fuzzy system are described. In subsection 3.2.3.2, the approaches to initialize the parameters for type-2 fuzzy systems are presented.

3.2.3.1 Identification of Parameters for Type-1 Fuzzy System

For construction of rule-base and membership functions to automate the rule-based fuzzy system, clustering based methods have been commonly used, in particular, for type-1 fuzzy systems [195], [196], [197], [9], [198]. Cluster analysis can be used to construct fuzzy rule-base and design membership functions. The clustering concept in relation to the rule-base extraction is briefly depicted in Fig. 3.8. The parameters of the MFs are obtained from the partitions. Each partition provides information such as centroid of a cluster, standard deviation of data objects within the cluster, all which can be easily used to derive membership functions.

As fuzzy sets are fully characterized by their membership functions, it is important to determine a set of appropriate membership functions for construction of a rule-based fuzzy logic system. Once the fuzzy sets have been established, the next step is to associate them with their membership functions. A membership function may come in many shapes such as triangular, trapezoidal, Gaussian, general bell, and sigmoidal. Some of membership functions that characterize fuzzy sets are widely used because of the ease of determining the parameters that specify them. It has been reported that the shapes of membership functions can effect the fuzzy inference in a rule-based fuzzy system [199] and the shape of if-part fuzzy set has been found to effect fuzzy logic systems that approximate continuous functions [200]. In addition to the shape of membership functions, values of the parameters used to design membership functions are equally important as they highly effect performance of the fuzzy logic systems.

The premise parameters of a rule-based fuzzy system are often non-linear in nature [201]. To ease the structure identification process, sample probability distributions were suggested in order to identify parameters of membership functions of input variables

using the centres of cluster-like regions [202]. On the other hand, it is a common practice to use an MF shape with a simpler representation and easier implementation [203].

The consequent part of a TSK fuzzy model is usually determined by the estimation of parameters of the linear regression models [201]. In order to find the consequent coefficient parameters defined in the linear regression model the least squares approach is commonly used. The linear regression model can be expressed as:

$$Y = LW \tag{3.20}$$

$$W = \begin{bmatrix} f_i^1 & f_i x_{i1}^1 & \cdots & f_i x_{ik}^1 \\ \vdots & \vdots & \vdots & \vdots \\ f_i^n & f_i x_{i1}^n & \cdots & f_i x_{ik}^n \end{bmatrix} \tag{3.21}$$

$$L = \begin{bmatrix} a_0^1 & a_1^1 & \cdots & a_k^1 & \cdots & a_0^n & a_1^n & \cdots & a_k^n \end{bmatrix} \tag{3.22}$$

where W is the weighted matrix of inputs and n is the number of input-output data pairs of the training data set; and L represents the unknown regression coefficients. The least squares method minimises the squared error E in order to approximate the linear function determined:

$$E = \sum_{i=1}^{n} (y_i - f(\vec{x_i}))^2 \tag{3.23}$$

where y_i and $f(\vec{x_i})$ are observed data and predicted data respectively, and n is the number of samples.

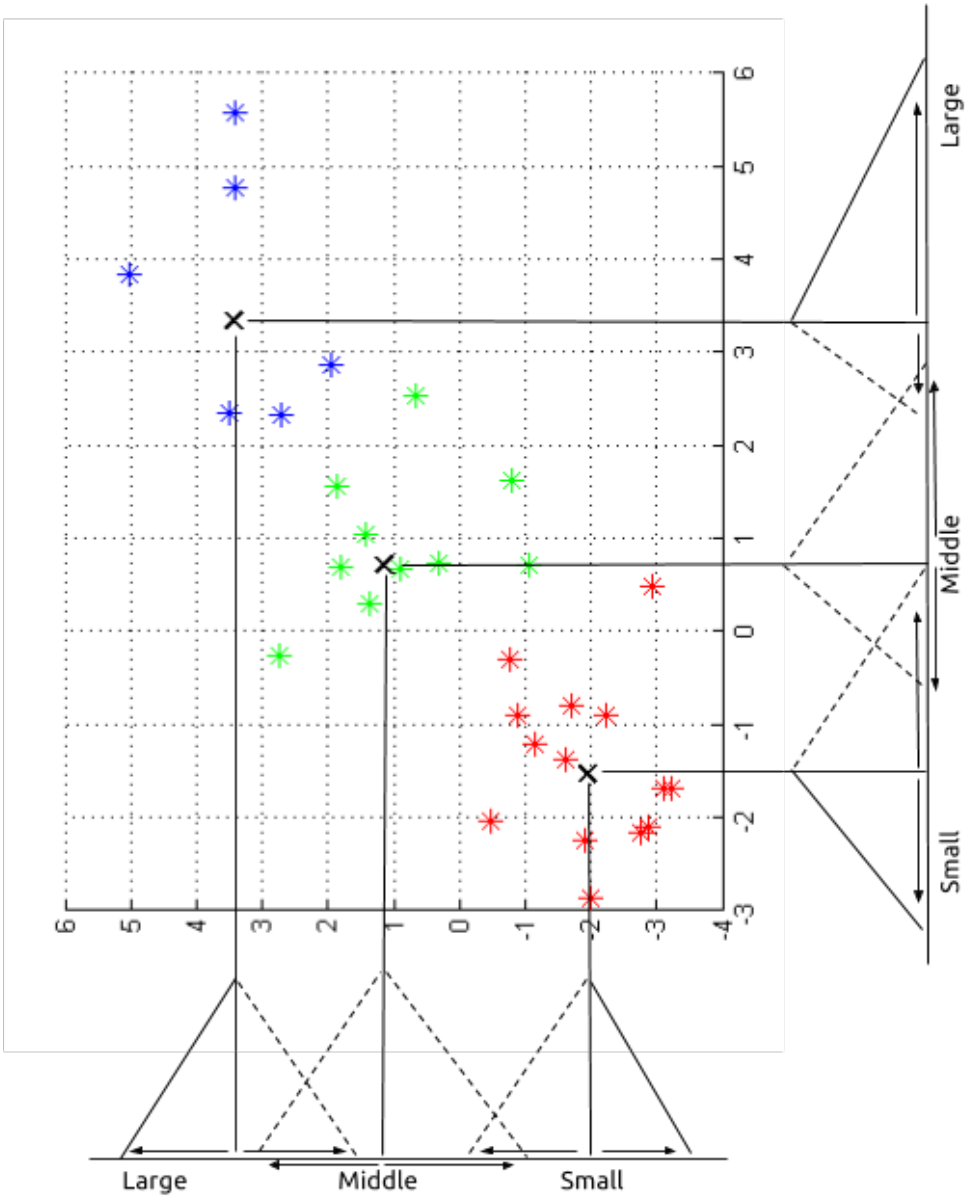


FIGURE 3.8: Illustration of determination of the initial values of the parameters of triangular membership functions using a cluster analysis.

3.2.3.2 Identification of Parameters for Type-2 Fuzzy System

To our best knowledge, there is no established method addressed in the literature to initialize the parameters of a type-2 fuzzy system. Common practice is the arbitrary initialization of these parameters then a learning method used in order to optimize them. The clustering approach that is used to identify the parameters for a type-1 fuzzy system can also be considered to be used in type-2 fuzzy rule-based systems. Moreover, improvements over existing clustering methods to be applicable to type-2 fuzzy system may be possible. Therefore, one aim of this research study is to develop a new clustering concept in order to identify the parameters of an interval type-2 fuzzy system. The clustering methods that will be employed in this research study are described and analysed in the consequent sections.

In a type-2 fuzzy logic system, the parameters of the membership functions are often need to be set. In the case of the Gaussian membership function, for an IT2-FS, the parameters of the membership functions can be uncertain. The primary membership functions for each antecedent IT2 fuzzy set may have uncertain means and fixed standard deviations or uncertain standard deviations and fixed means as depicted in Fig. 3.9 and Fig. 3.10, respectively [193]. A Gaussian type-1 fuzzy set can be characterized by the

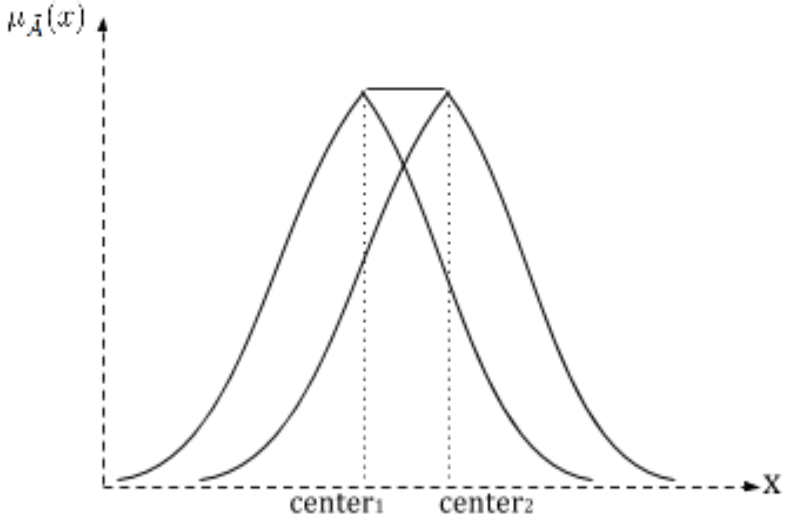


FIGURE 3.9: Gaussian membership function with fixed standard deviation and uncertain means.

following membership function:

$$\mu(x) = e^{-\frac{(x-c)^2}{2(\sigma)^2}} \tag{3.24}$$

where $\mu(x)$ is the degree of membership for input variable x; and c and σ are the centre and standard deviation that characterizes the Gaussian type-1 fuzzy set, respectively. The bounded region of a Gaussian interval type-2 fuzzy set is often formed by the blurring of mean or standard deviation of a Gaussian type-1 membership function [204]. In the case of blurring the mean to form an interval $[c_1, c_2]$, the UMF can be expressed as:

$$\overline{\mu}(\mathrm{x}) = \begin{cases} e^{-\frac{(x-c_1)^2}{2(\sigma)^2}}, & x < c_1 \\ 1, & c_1 \leq x \leq c_2 \\ e^{-\frac{(x-c_2)^2}{2(\sigma)^2}}, & x > c_2 \end{cases} \tag{3.25}$$

and the LMF can be expressed as:

$$\underline{\mu}(x) = \min(e^{-\frac{(x-c_1)^2}{2(\sigma)^2}},\ e^{-\frac{(x-c_2)^2}{2(\sigma)^2}}) \tag{3.26}$$

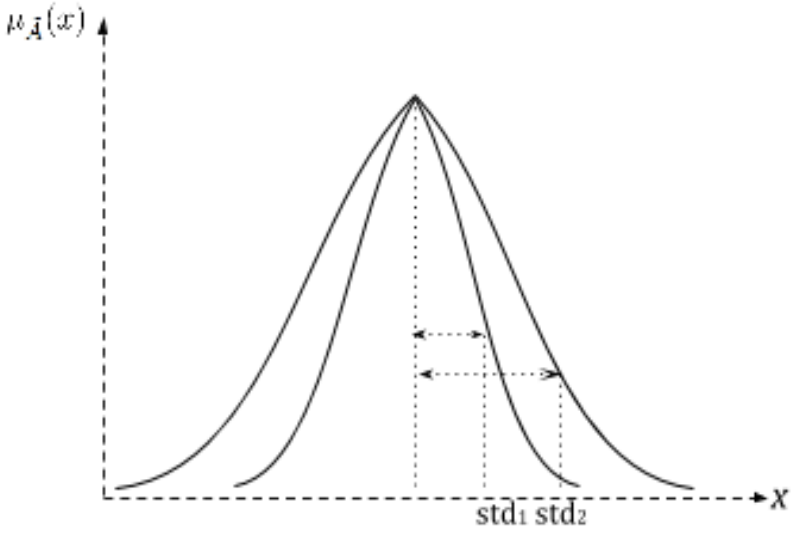


FIGURE 3.10: Gaussian membership function with fixed mean and uncertain standard deviations.

In the case of blurring the standard deviation to form an interval $[\sigma_1, \sigma_2]$, the UMF can be expressed as:

$$\overline{\mu}(x) = e^{-\frac{(x-c)^2}{2(\sigma_2)^2}} \tag{3.27}$$

and the LMF can be expressed as:

$$\underline{\mu}(x) = e^{-\frac{(x-c)^2}{2(\sigma_1)^2}} \tag{3.28}$$

In addition to arbitrary approach, there exists a few alternative methods suggested in the literature for adjusting the bounded region for the interval type-2 fuzzy sets to represent the uncertainty. In the work of Tan et al [205], once the upper MF is determined, the footprints of uncertainty associated with the interval type-2 membership functions are formed by varying the parameters of the lower MF. They suggested two strategies to select the FOU associated to the MFs which are illustrated in Fig. 3.11 and Fig. 3.12. The former strategy adjusts the FOU by varying the height of the lower MFs. The latter strategy adjusts the height as well as the left and right end points of the lower MFs.

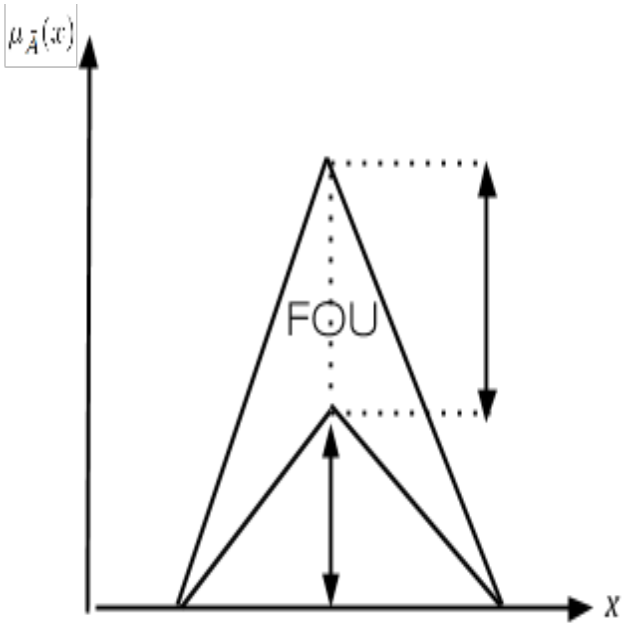


FIGURE 3.11: FOU design by varying the height of the lower MF.

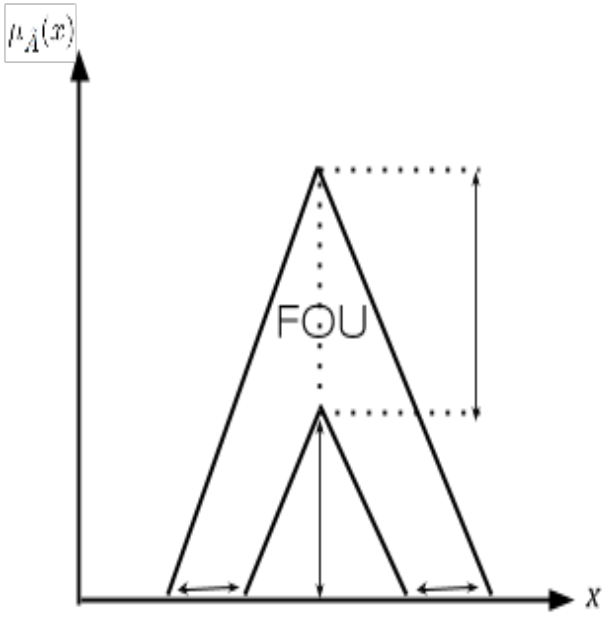


FIGURE 3.12: FOU design by adjusting the height, left and right-points of the lower MF.

3.2.4 Optimisation of Fuzzy Logic Systems

The lack of learning capability of fuzzy systems generated a research interest on learning approaches to determine optimum values of the parameters of fuzzy logic systems including membership functions. As previously discussed in Section 3.2.3 the construction of a rule-based fuzzy system and its membership functions can be automatized. The research showed that, particularly in the last two decades, fuzzy systems can be enhanced with learning and adaptation capabilities [8]. Neural and genetic fuzzy systems are the two such approaches that augment fuzzy systems with learning and adaptation methods. Neuro-fuzzy systems, the combination of neural networks and fuzzy logic, use a machine learning algorithm to determine the parameters of a fuzzy rule-based system by processing data samples [206], [207], [208]. The Adaptive Neuro Fuzzy Inference System (ANFIS), introduced in [209], is one of the most successful examples of neuro-fuzzy systems and presents the architecture and learning principle of the adaptive networks. Genetic-fuzzy systems, which combine the genetic algorithms [210] and fuzzy logic, employ an evolutionary learning process to automate the design of the rule-based fuzzy system based on the search capability of the genetic algorithms [211], [212], [213], [214], [215]. A different approach to hybridisation is the use of simulated annealing [216], [217]. This is basically a fuzzy system augmented by an optimization process based on a simulated annealing algorithm.

3.3 Support Vector Regression (SVR)

The support vector approach is based on the statistical learning theory (also known as VC theory) which was introduced in the sixties [218]. The statistical learning theory, aiming at estimation of a function from the given data set, remained theoretical until the nineties. In the mid nineties, Support Vector Machine (SVM) learning algorithm was proposed based on this theory, leading the theory becoming in practice [219], [220]. SVMs search for an optimal separating hyperplane from a given collection of data. Data samples are mapped to a high-dimensional feature space so that they can be separable with a linear hyperplane. As the mapping is non-linear, an adequate kernel function has to be chosen. Therefore, two classes that are separated with a maximized margin from each other, are revealed.

SVMs not only can be used for classification but also for real-value estimation tasks as well. The regression form of SVM is SVR [221] and has been shown to have superior performance in many applications [222], [223], [224]. SVR uses the ϵ-insensitive loss function as depicted graphically in Fig. 3.13 (figure adapted from [225]). One advantage of using this function is that it can tolerate against noise. SVR approximates a linear function $f(x)$ in the following form:

$$f(x) = w^T x + b \tag{3.29}$$

where the coefficients w and b are the weight vector and bias term, respectively. This linear function can be constrained to the following optimisation problem:

$$\min \frac{1}{2}\|w\|^2 + C\sum(\xi_+ + \xi_-) \tag{3.30}$$

where ξ^+, ξ^- are the two nonzero slack variables in both directions. The bounded area aims at fitting the data with an admissible parameter ϵ. The constant parameter $C > 0$ is the trade-off that it optimizes (3.30) between the complexity (flatness) of the function and toleration up to the distance value of data samples outside the bounded region (slack variables) which deviate greater than ϵ. The data samples that are outside of the bounded zone within the distance of slack variables are the support vectors. The minimisation function (3.30) is subject to:

$$\begin{aligned} &y - (w^T x + b) \leq \epsilon + \xi_+ \\ &(w^T x + b) - y \leq \epsilon + \xi_- \\ &(\xi_+, \xi_-) \geq 0 \end{aligned} \tag{3.31}$$

The constrained optimisation problem (3.30) and (3.31) can be solved with the method of standard dualization. Dual formulation reformulates the optimisation function using the Lagrange multipliers with the help of a dual set of parameters. After a set of steps (for details see [226]), dual optimisation problem yields to the following solution:

$$f(x) = \sum_{i=1}^{n}(\alpha - \alpha^*)K(x_i, x) + b \tag{3.32}$$

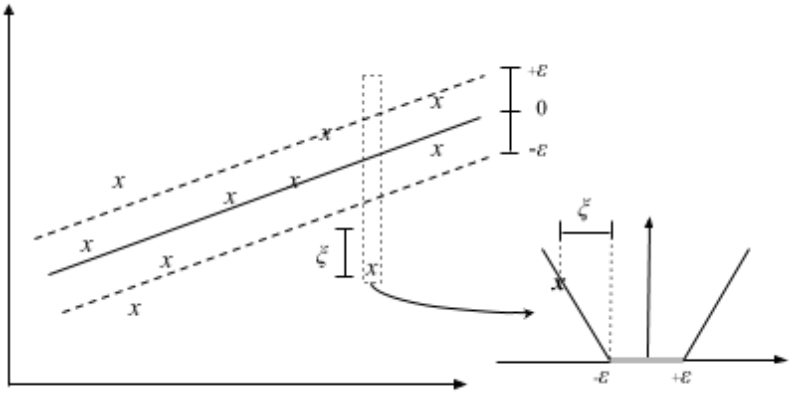


FIGURE 3.13: ϵ-insensitive loss function for a linear SVM.

where α and α^* are Lagrange multipliers; and the kernel function is represented by $K(x_i, x)$.

The kernel function can map the non-linear input space to the high-dimensional feature space so that a linear solution may be possible. One common problem in the support vector based approach is that it is not easy to determine which kernel function can be used [227]. The choice of a kernel function may depend on several factors, particularly depends on the data set that is being used. Once the kernel function is determined, the parameters C, ϵ and the kernel parameter (depending on the chosen kernel function) are required to be set properly. Hence, a proper set of parameters can lead a suitable SVR solution that can best model the data set in use. Once the parameters are selected properly, one can expect a better generalization performance from the constructed SVR model.

Another common limitation of the support vector approach is its efficiency for very large data sets. It can be very hard to train such data sets as of the availability of the millions of support vectors. The training for very large data sets as well as the fixing of kernel function, still remain open research issues.

3.4 Revealing Clusters in Feature Space

The clustering is an exploratory data analysis method that groups objects into sets having similar characteristics. Cluster analysis helps pre-process the data for an additional analysis, arrange and determine the characteristic prototypes of the data, identify closely

connected regions of data, and visualize the data [228]. As clustering is unsupervised method, it is different than the classification. Given a data set D = x_1, x_2, x_3, ..., x_n in X, the objective is to learn a function f : X $\rightarrow$ c_1, ..., c_k that each cluster c in c_1, ..., c_k is formed through placing each object x_i to its closest group. The function f maps X to a feature space H as in the form of f : X $\rightarrow$ H. Therefore objects within the same group have a higher cluster similarity than the objects in different groups.

Although the clustering concept has been studied for many years, there does not seem to be a definite taxonomy of the clustering methods. Several taxonomies, most of which are common, have been given in the literature [229], [230], [231]. In general, clustering methods can be classified into two main types. They are hierarchical clustering and partitional clustering [232]. Hierarchical clustering methods organise data into a nested sequence of partitions and provide a graphical representation called dendrogram. As opposed to the nested sequence, partitioning clustering methods provide separate clusters for each group of objects in the data [233].

3.4.1 K-means Clustering

K-means (also known as Hard c-Means [234]) clustering is one of the basic methods in clustering [235]. It begins with arbitrarily set initial cluster centres. Then in each iteration the nearest cluster for each object is computed and the object is assigned to the nearest cluster. After all objects are assigned to the clusters, new cluster centres are computed. This process continues until a stopping criteria (e.g., mean squared distance) is satisfied.

There is no simple and generally good method for determining the number of clusters and the initial placement of centers [232], [236]. The cluster centres converge sensitive to different initial points [237]. A general strategy for the method of initialization is to run the algorithm with random initial centres [232]. The initial centers may also be chosen by taking a random sample of data points [238]. There are number of variants of k-means algorithm, due to its simplicity and flexibity, Lloyd's algorithm is widely used [239].

3.4.2 Fuzzy c-Means Clustering

Fuzzy c-Means clustering is an expanded version of the k-means clustering and useful in analysing data sets in which the boundaries among the clusters are uncertain. In the nature of fuzzy logic, each point has a degree of membership to clusters rather than belonging to only one cluster. The concept of fuzzy c-partitions was first introduced by Ruspini [240] and then followed by Bezdek who developed fuzzy c-Means clustering [241], [242]. FCM partitions the data set into various clusters by assigning a degree of membership for each data object to all the clusters. The FCM algorithm, introduced by Bezdek [241], is one of the widely used methods in fuzzy clustering. Rather than assigning each data object into only one cluster as in the k-means, fuzzy c-Means relaxes this crisp approach by giving more degrees of freedom to the data object in the data set by ensuring that the data object belongs to all the clusters with varying degrees of membership. The clustering process iteratively calculates cluster centres and degrees of memberships of each data point until an objective function is satisfied. The FCM algorithm can be summarized as follows.

The fuzzy c-means clustering model attempts to obtain partitions (V) for the unlabeled object data in R^p. The data $X = x_1, x_2, \ldots, x_n$ represents the data objects where each data object is a vector in R^p. The fuzzy c-partition of the data set $U = [u_{ij}]$ is a $c \times n$ membership matrix where u_{ij} is the degree of membership of the j^{th} sample for the i^{th} cluster; n is the number of samples and c is the number of clusters. Then, the sum of membership values of an object should be equal to one.

$$\sum_{i=1}^{c} u_{ij} = 1, \text{where } \forall j = 1, 2, ..., n \tag{3.33}$$

A distance measure d_{ij} can defined by

$$d_{ij} = \|x_j - c_i\| \tag{3.34}$$

Measuring the distances between data objects and cluster centers in any inner product norm, and a membership of data objects with a weight exponent minimizes the objective

function, J_m,

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{\tau} {d_{ij}}^2,\ \tau \in [1,\infty) \tag{3.35}$$

where τ is regarded as fuzzification factor. The cluster centers can be updated by using the membership degrees as given in (3.36)

$$\mathrm{c_i} = \frac{\sum_{j=1}^{n} u_{ij}^{\tau} \mathrm{x_j}}{\sum_{j=1}^{n} u_{ij}^{\tau}} \tag{3.36}$$

The membership values of each data object can then be found by using the following mathematical expression

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{\tau-1}}} \tag{3.37}$$

The process continues until the objective function is minimized or the number of iterations reaches a preset value.

FCM has been further analysed, improved, and applied and many variations of the algorithm have been developed. Nascimento et al proposed a model, named fuzzy clustering multiple prototype, that defines the underlying fuzzy c-partition in such a way that the membership of an object to a cluster expresses a part of the cluster's prototype reflected in the object [243]. FCM is extended to include data sets whose feature values are continuous random variables [244]. Furthermore, FCM with the added possibilistic approach may suggest more accurate results [245], [246], [247], [248], and dynamic FCM addresses and analyses the dynamic data environments [249].

In relational data clustering, object-data is not available and the clustering process is performed based on a similarity/dissimilarity relational data. One of the first examples

of fuzzy relational clustering is proposed by Roubens [250] as in the form of fuzzy non-metric model (FNM). This model assumes a dissimilarity relation R satisfying three constraints: $r_{ij} \geqslant 0$, $r_{ii} = 0$ and $r_{ij} = r_{ji}$. Hathaway and Bezdek reformulated the optimisation function J_m [251]. The reformulation of the optimisation function K_m eliminates the use of protoype means. K_m takes a form which is dual of J_m when the pairwise distances of object-data define the relation matrix R.

$$K_m = \sum(\sum\sum(u_{ij}^{\tau}u_{ik}^{\tau}\|x_j - x_k\|^2)/(2\sum u_{it}^{\tau})) \tag{3.38}$$

The optimisation function can be redefined as

$$K_m = \sum(\sum\sum(u_{ij}^{\tau}u_{ik}^{\tau}r_{kj})/(2\sum u_{it}^{\tau})) \tag{3.39}$$

where $r_{kj} = \|x_j - x_k\|^2$.

The relational fuzzy c-means (RFCM) clustering model attempts to obtain partitions for the relational data D=$[D_{ij}]$ where D consists of distances some data set X. The number of clusters is fixed to c, where $2 \leq$ c $\leq$ n. The fuzzification factor should be τ>1 and partition matrix $U^0 \in M_{fcn}$ is initialised.

The c-mean vectors $v_i = {v_i}^t$ can be updated by using the membership degrees $U = U^t$, for $1 \leq$ i $\leq$ c:

$$v_i = \frac{(U_{i1}^{\tau},\ \ldots\ , U_{in}^{\tau})}{(U_{i1}^{\tau} + \ldots + U_{in}^{\tau})} \tag{3.40}$$

and then calculate d_{ik} in (3.41) for $1 \leq$ i $\leq$ c and $1 \leq$ k $\leq$ n,

$$d_{ik} = (Rv_i)_k - (v_i^T Rv_i)/2 \tag{3.41}$$

The partition matrix U^t is updated to $U = U^{t+1} \in M_{fcn}$ that satisfies (3.42) and (3.43), for each k=1, ..., n. If d_{ik}>0 for all i, then (3.42) otherwise (3.43) that means at least

one $d_{ik} = 0$.

$$U_{ik} = \frac{1}{(\frac{d_{ik}}{d_{1k}} + \frac{d_{ik}}{d_{2k}} + ... + \frac{d_{ik}}{d_{ck}})^{1/\tau - 1}} \tag{3.42}$$

$$U_{ik} > 0 \text{ if } d_{ik} = 0,\ U_{ik} \in [0, 1],$$
$$\text{and } (U_{ik} + ... + U_{ck}) = 1. \tag{3.43}$$

The process continues until the objective function K_m is minimized (3.38 or 3.39) or the number of iterations reaches a preset value.

3.4.3 Hierarchical Clustering

Hierarchical clustering algorithms organise data into a cluster tree or dendrogram [252], [253]. The cluster tree is a multi-level hierarchy and set of clusters are obtained by cutting this cluster tree at a predefined level of the hierarchy. Generally, hierarchical clustering algorithms can be divided in two main types. They are agglomerative clustering methods and divisive clustering methods [232].

Agglomerative clustering algorithms are bottom-up type of hierarchical clustering algorithms [254], [255]. It begins by finding proximity of each object relative to each other object in the data set. The objects that are closer to each other are linked to binary clusters. Then newly formed clusters are linked into larger binary clusters. This process continues until all the data objects are grouped under the root node in the form of a hierarchical cluster tree. Ultimately, the set of clusters are obtained by cutting the dendrogram at the predefined level and all objects in the data set are assigned to clusters determined at this level of the hierarchical tree. This process is depicted in Fig. 3.14 where an example dendrogram groups 9 data objects into clusters at different levels [256]. The lines show the levels and the data objects in the same branch of the dendrogram below the line are grouped into clusters at that level.

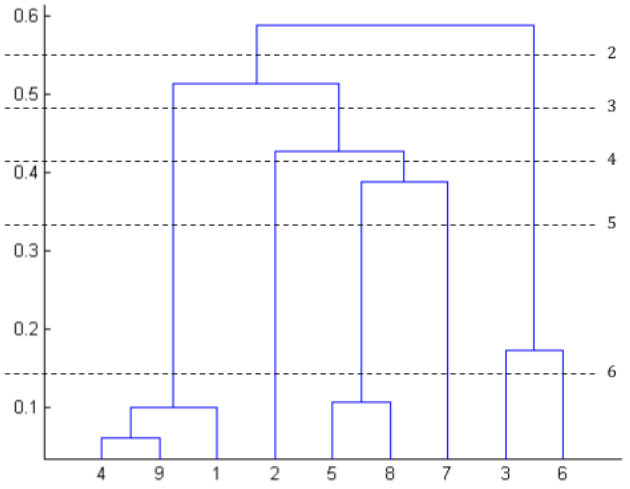


FIGURE 3.14: An example dendrogram.

Divisive clustering is a top-down type of hierarchical clustering and it moves in the opposite way. It begins with a root cluster that the entire data set belongs to, and it progresses by dividing clusters into two in each level. This process continues until leaf clusters, each of which contains one data point, are obtained. Clustering n data points in a data set requires $2^{(n-1)-1}$ possible binary divisions [229]. In divisive clustering, the computational cost is high, thus in practice it has no wide use as compared with agglomerative clustering. A further discussion on divisive clustering algorithms and their applications can be found in [229], [257].

It should be noted that hierarchical clustering can be sensitive to dimensionality as the number of dimensions increase [258]. A fixed number of data samples might become sparse in the high-dimensional feature space. The difference in distance or similarity between the nearest and farthest data samples becomes relatively uniform or approaches zero as the dimensionality increases [259].

3.4.4 Determining the Number of Clusters

Another concern in clustering is the quality of the partitions. Consequently, cluster validity and ensemble methods should be considered in order to improve clustering results. Determining optimum number of clusters is also an important part of the cluster validity. Cluster validity is affected by the parameters in order to find out correct number of clusters and the validation of the clusters that the data is partitioned into

[260]. Additionally, a new model is proposed in [261] that employs validity indexing part to determine the number of clusters for several clustering methods.

The approaches for revealing the number of clusters [262], [263], [237] are based on the use of cluster validity indices in line with optimization methods such as particular swarm optimization [264] and genetic algorithms [210]. Dunn's index [265] is a common validity index that is employed in interpretation and validation of the number of clusters for the provided data set.

Visual assessment of cluster tendency (VAT) is a method to visually assess the cluster tendency of a given data set [266]. The data set can be represented either as object vectors or by numerical pairwise dissimilarity values. The objects in the data set are reordered in the form of a matrix. The pairwise similarities/dissimilarities of data objects are displayed as an intensity image. By observing visually darker blocks of the reordered matrix laying on the diagonal, the number of clusters that would be in the analysed data set is revealed. The improved VAT (iVAT) algorithm has been shown to overcome the problems (e.g., lack of showing the cluster tendency) of VAT for some tough cases [267]. Clustering ordered dissimilarity data algorithm (CLODD) can cluster either object or relational data and suggests clusters in the reordered relational data by recognizing the blocky structure in the reordered data [237].

A cluster silhouette is another kind of method that helps determining the natural number of clusters of data. This method represented as a graph and the interpretation of this plot provides an insight information about how tightly the samples in a data set are grouped into their respective partitions. The equation is given as follows:

$$s_i = \frac{b(i) - a(i)}{max(b(i), a(i))} \tag{3.44}$$

where a(i) is the average dissimilarity of i to each of other samples in the same partition and b(i) is the lowest average dissimilarity of i to a partition other than the which it is assigned.

The distance measure to be used in order to find the value of the dissimilarity can be any measure. The value of s(i) always is in the interval between -1 and 1. If the value of s(i) is closer to 1, it means the data sample is appropriately clustered. On the other hand,

if the value of s(i) is closer to -1, this means that the data sample is poorly clustered. In this case, the neighbour partition may be a better option that if the sample was assigned to it. The smaller values of a(i) indicates that a better grouping for the sample i is decided. The value of b(i) often indicated a neighbour partition as it is the most likely partition the sample can be assigned to other than its existing cluster.

3.5 Feature Selection Method

The bioinformatics data sets become challenging nowadays due to the rapid growth in their number of samples and features. Thus, a significant increase in processing time as well as space requirements is unavoidable. However, computational methods are mostly designed to work out low dimensional spaces. As a consequence, such data sets are increasingly computationally unmanageable and intractable in high-dimensional spaces where thousands or even ten-thousands features are available. Therefore, feature selection or feature reduction are commonly used to address the computational complexity of such data sets aiming at improving the performance of the computational models.

Feature selection is the process of selecting a set of features that improves the efficiency of the model [158], [268], [269], [270], [271]. Four key steps are involved in a typical feature selection process as shown in Fig. 3.15 (figure adapted from [269]). These are subset generation, subset evaluation, stopping criterion, and the validation of result.

Feature selection methods appear in many applications as a preliminary stage during the model building process. They can cope with large size features and help to eliminate those of the features which are irrelevant. They also aid in simplification of the model and address the curse of dimensionality problem. There are three main characteristics of feature selection methods [158] as shown in Fig. 3.16. They are: a) to improve the performance of the model, b) to provide a computationally efficient model, c) to present a new representation for the data set to be simpler to understand. As a consequence, a more generalized and interpreted model from the data can be obtained. It should be noted that, as the accuracy of results takes the centre stage in bioinformatics, computational efficiency is less important in bioinformatics research studies.

Zhao et al [268], proposed a repository for various feature selection methods and in this repository these methods are organized into three main categories of which are filter,

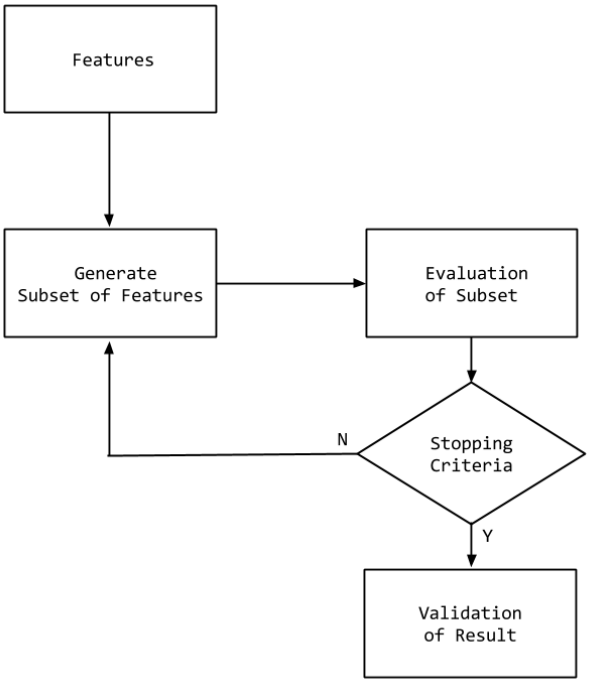


FIGURE 3.15: Key steps of feature selection.

wrapper and embedded models. Furthermore, they also categorized feature selection methods differently based on their characteristics. Some of these categories can be: 1) supervised or unsupervised, 2) univariate or multivariate, 3) variable ranking or subset selection. Somol et al [270] added the hybrid approach to the three main categories which aims for combining advantages of at least two of these aforementioned categories.

In this research study, three predictive models are used (SVR, Type-1 Fuzzy System,

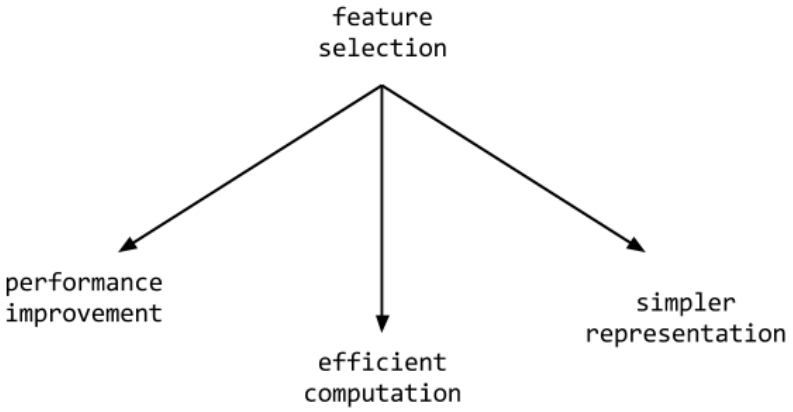


FIGURE 3.16: Characteristics of feature selection.

Type-2 Fuzzy System) on different kinds of peptide binding affinity data sets. Therefore, the feature selection method needs to be independent of any predictive model as they are required to be tested on unseen data to evaluate their performance. There are feature selection methods that do not require the output in their selection process. They are regarded as unsupervised feature selection methods such as Unsupervised Feature Selection Using Feature Similarity [272], Multi Cluster Feature Selection (MCFS) [273], Laplacian Score (LS) [274], Q-alpha [275].

Among all these methods MCFS has shown to present better results than other methods such as LS [276]. Therefore, MCFS is chosen to be used as a feature selection method in the preprocessing stage of the proposed predictive models. MCFS is an unsupervised feature selection method and uses information contained in eigenvectors by solving the generalised eigen-problem to preserve the multi-cluster structure of the data. This feature selection method finds a subset of features that can cope with any clustering structure within the data. The correlation of features between each other are assessed using spectral analysis without the need of any output or target label.

3.6 Performance Measurements of the Prediction Models

The quantitative measure for a peptide binding affinity is given as pIC50 (-log IC50) value for the peptide binding affinity data sets used in this research study. IC50 scale is the half maximal (%50) inhibitory concentration indicating the quantity of a substance required to inhibit a biological activity by half. In pharmaceutical biology, the IC50 scale is used for measuring the antagonist drug potency [277], [278]. It is often practice to convert IC50 scale to a pIC50 scale in molecular modelling studies [279], [280]. The high pIC50 scales indicate high potency whereas low pIC50 scales indicate low potency [281], [282].

There are different measurements used to assess capability of the predictive models. However, in order to maintain consistency over the published results and perform consistent comparison, the following measures; coefficient of determination (q^2) and spearman rank correlation coefficient (ρ) are used for the CoEPrA peptide binding affinity data sets. For the mouse class I MHC alleles, coefficient of determination (q^2) and average residual (AR) are used.

The measure q^2 is a statistical model based upon the proportion of variability in a data set [283]. When q^2 is close to 1 it suggests a model that has been successfully constructed. Negative q^2 values indicate that model poorly approximates the expected values. q^2 can be expressed as:

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_{exp} - y_{prd})^2}{\sum_{i=1}^{n}(y_{exp} - \overline{y}_{exp})^2} \tag{3.45}$$

where y_{exp} and y_{prd} are the expected and predicted values of the peptide binding affinity, respectively, n is the number of peptides and $\overline{y}_{exp}$ is the mean of all expected values in the prediction data set.

The spearman rank correlation coefficient (ρ) [284] is used to measure the statistical dependence between two variables. The value of ρ ranges between +1 and -1 showing perfect correlation at each end.

$$\rho = 1 - \frac{6\sum(y_{exp} - y_{prd})^2}{n(n^2 - 1)} \tag{3.46}$$

where y_{exp} and y_{prd} are the expected and predicted values of the peptide binding affinity, respectively, n is the number of peptides in the prediction data set.

The average residual measure is another metric that is used particularly in experimenting models for the mouse class I MHC alleles. AR can be expressed as:

$$\mathrm{AR} = \frac{\sum_{i=1}^{n}|y_{exp} - y_{prd}|}{n} \tag{3.47}$$

where n is the number of peptides in the allele. A successful prediction can be achieved with lower values of AR whereas its higher values show poorer predictions.

Improvement gain or loss of one method ($\text{Model}_{\text{new}}$) over another ($\text{Model}_{\text{old}}$) is used to show the performance of the proposed models.

$$\%\text{I}_{\text{gain/loss}} = \frac{\text{Model}_{\text{new}} - \text{Model}_{\text{old}}}{|\text{Model}_{\text{old}}|} \times 100\% \tag{3.48}$$

In addition, overall improvement gain or loss of a group of models is computed as follows:

$$\%\text{Overall}_{\text{gain/loss}} = \frac{\sum_{i=1}^{n} \%\text{I}^{i}_{\text{gain/loss}}}{n} \tag{3.49}$$

where n is the number of models in the group.

Chapter 4

Description and Selection of Amino Acids based Features for Peptide Binding Affinity Prediction

4.1 Introduction

Understanding of the peptide data sets is important as they are used to find a solution for the peptide binding affinity problem that is dealt with using the predictive modelling. Therefore, peptide data sets and how they are encoded into their features are clarified in this chapter before presenting SVR-based fuzzy systems to quantitatively predict binding affinities between MHC proteins and peptides in Chapters 5 and 6. In Section 4.2, materials and methods are explained. Characteristics of two groups of data sets are presented as they are used to demonstrate the ability of the proposed predictive models to generalise for the unseen peptides. The amino acid based features which are used to encode the feature space, are presented. Section 4.3 is the results and discussion section which presents the selection of amino acids based features from this feature space. Finally, chapter is concluded in Section 4.4.

4.2 Materials and Methods

In this section the characteristics of two groups of data sets are presented. First group of data sets are the CoEPrA peptide binding affinity data sets that are formed of four tasks which is detailed in Section 4.2.1. Each task has separate train and test data sets. Each data set consists of peptide samples along with their attributes; peptide no, peptide residue, and expected real-value binding affinity of peptide. These data sets are made available in Appendix C. Second group of data sets are the mouse class I MHC peptide binding affinity data sets (H2-Db, H2-Kb and H2-Kk) and explained in Section 4.2.2. Entire data set is provided for each of the mouse class I MHC peptide allele. Each data set consists of epitope samples along with their attributes; epitope no, epitope residue, and expected real-value binding affinity of epitope. These data sets are made available in Appendix D.

4.2.1 CoEPrA Peptide Binding Affinity Data Sets

The publicly available high-dimensional peptide data sets provided at the Comparative Evaluation of Prediction Algorithms (CoEPrA) modeling competition [285] are used in this research study. The summary of these data sets are provided in Table 4.1 and Table 4.2.

Amino acid occurrences in training and testing peptide data sets for each experiment are given in Table 4.3 - Table 4.6. In these data sets physico-chemical descriptors have been provided for each peptide (for both calibration and prediction data sets). Each amino acid in a peptide is described by 643 descriptors. Task 2 consists of octa-peptides that have a total of 5144 (643x8 = 5144) descriptors. All other tasks have nona-peptides that have a total of 5787 (643x9 = 5787) descriptors. The task (for all tasks except Task 4) is to predict actual affinity values (pIC50) for peptides from the amino acid descriptors. For Task 4 it is clear that the expected values are not given as pIC50 values. But it cannot be determined which measure it is, as it is not provided on the aforementioned website. For this reason the performance of the model for Task 4 is more likely based on the prediction of correlation rather than the actual values. The statistics (range, mean and standard deviation) of the binding affinities of the peptides of each task are given in Table 4.2.

Table 4.3 - Table 4.6 show the distribution of amino acids placed on the peptide locations for each of the calibration and prediction data sets of related tasks. Data set analysis of Task 1 shows some strong preferences on various peptide locations. Proline (P) at position 4 and 6 and Valine (V) at position 9 contributes strongly on the Task 1 data sets. Although Leucine (L) at position 2 contributes weakly on the Task 1 model, prediction data set contains Leucine (L) at position 2 strongly, which in turn makes the prediction of Task 1 is rather difficult. For the Task 2, 76 octomer peptides were used to train the model using the calibration data set. Every anchor location for the octomer data sets (Task 2) have one particular binding position. The amino acids with high occupancy rate are Phenylalanine (F), Glutomic Acid (E), Serine (S), Threonine (T), Glycine (G), Asparaigne (N), Leucine (L), Isoleucine (I) with approximately 60 occurancies at separate respective positions. Tasks 3 and 4 use the same calibration data set with different prediction data sets. Leucine (L) at position 2 and Valine (V) at position 9 strongly contributes on the Task 3 model. However Task 4 prediction data set differs from Task 3 prediction data set with rather low occupancy rate for Leucine (L) and Valine (V).

TABLE 4.1: General characteristics of the peptide data sets used for the prediction of peptide binding affinity.

Data Sets	Number of Peptide Sequences		Nature of Peptide	Number of Descriptors
	Training	Testing		
Task 1	89	88	nona-peptide	5787
Task 2	76	76	octa-peptide	5144
Task 3	133	133	nona-peptide	5787
Task 4	133	47	nona-peptide	5787

TABLE 4.2: The statistics of the binding affinity of peptides for each peptide data set.

Data Sets	Training				Testing			
	Min	Max	Mean	Std	Min	Max	Mean	Std
Task 1	2.94	8.65	5.41	1.01	3.13	8.17	5.41	0.95
Task 2	5.01	8.34	7.55	0.77	5.01	8.40	7.58	0.74
Task 3	4.30	8.77	7.08	0.82	5.08	8.96	7.10	0.80
Task 4	4.30	8.77	7.08	0.82	13.0	121.0	61.0	34.0

TABLE 4.3: Amino acid occurrences in training and testing nona-peptide data sets for CoEPrA Peptide Binding Affinity Task 1.

Training

Amino Acid	**Location**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	1	2	2	0	0	0	1	2	14
Arginine	5	0	0	0	0	0	0	0	0
Asparagine	1	0	6	1	0	1	1	11	0
Aspartic acid	0	0	29	4	0	2	1	2	1
Cysteine	1	1	2	1	0	1	1	2	0
Glutamine	0	0	1	10	4	2	2	3	0
Glutamic acid	0	0	0	0	0	0	2	3	0
Glycine	3	0	1	6	16	1	1	1	2
Histidine	1	1	3	1	1	0	8	1	1
Isoleucine	3	2	3	0	4	1	2	1	5
Leucine	3	6	5	2	10	1	1	4	6
Lysine	2	0	1	2	0	0	0	0	1
Methionine	1	4	4	0	1	1	0	0	0
Phenylalanine	9	1	13	1	33	2	11	0	1
Proline	1	1	0	52	1	50	14	4	1
Serine	2	0	3	4	1	3	4	12	1
Threonine	0	7	1	3	5	6	1	39	3
Tryptophan	0	0	12	0	1	0	1	2	1
Tyrosine	2	1	3	0	3	14	1	1	1
Valine	3	1	0	2	9	4	37	1	51

Testing

Amino Acid	**Location**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	3	0	4	1	1	1	5	2	13
Arginine	4	0	0	3	3	1	0	1	0
Asparagine	2	1	3	1	0	3	0	5	1
Aspartic acid	0	1	25	8	2	0	1	5	0
Cysteine	0	1	1	0	1	2	1	2	2
Glutamine	0	2	0	11	0	1	0	2	1
Glutamic acid	0	0	2	3	2	0	1	5	1
Glycine	3	1	3	1	16	2	1	4	0
Histidine	2	0	1	1	6	1	11	2	0
Isoleucine	29	4	2	1	6	4	3	4	6
Leucine	3	65	6	0	8	2	6	4	16
Lysine	2	0	3	0	0	0	0	0	0
Methionine	1	3	1	0	0	1	3	1	1
Phenylalanine	8	0	17	1	24	5	8	2	0
Proline	0	0	2	45	2	46	10	1	0
Serine	4	1	2	4	1	2	3	8	0
Threonine	3	5	2	4	0	3	2	39	1
Tryptophan	2	1	10	2	2	0	0	1	0
Tyrosine	19	0	3	1	5	10	1	0	0
Valine	3	3	1	1	9	4	32	0	46

TABLE 4.4: Amino acid occurrences in training and testing octa-peptide data sets for CoEPrA Peptide Binding Affinity Task 2.

Training

Amino Acid	Location							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	1	0	1	1	1	0	0	1
Arginine	0	0	1	1	0	1	0	1
Asparagine	2	0	1	0	2	66	1	9
Aspartic acid	1	1	1	1	0	1	2	1
Cysteine	0	0	0	0	0	0	1	0
Glutamine	2	1	1	0	0	0	1	1
Glutamic acid	0	67	0	1	0	1	2	0
Glycine	1	2	1	1	65	2	0	1
Histidine	1	0	1	0	0	0	1	1
Isoleucine	1	1	1	1	1	0	1	57
Leucine	1	1	2	1	1	0	64	0
Lysine	1	1	1	1	0	3	1	0
Methionine	1	0	0	0	0	0	0	1
Phenylalanine	60	1	2	1	1	0	1	0
Proline	1	0	0	1	1	0	1	0
Serine	1	0	63	1	1	0	0	1
Threonine	0	1	0	61	0	0	0	0
Tryptophan	1	0	0	1	1	1	0	1
Tyrosine	0	0	0	1	0	0	0	0
Valine	1	0	0	2	2	1	0	1

Testing

Amino Acid	Location							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	1	4	0	0	1	1	1	0
Arginine	1	0	0	0	1	0	1	0
Asparagine	0	1	0	1	0	59	0	10
Aspartic acid	1	0	0	0	1	0	0	0
Cysteine	0	0	0	0	0	0	0	0
Glutamine	0	0	0	1	1	1	0	0
Glutamic acid	1	62	1	0	1	1	0	0
Glycine	0	0	0	1	63	1	1	0
Histidine	1	1	1	2	1	1	0	0
Isoleucine	0	0	2	0	0	1	1	55
Leucine	1	1	0	1	0	2	64	2
Lysine	0	0	1	1	1	0	0	1
Methionine	0	1	1	1	1	2	1	0
Phenylalanine	68	0	2	0	1	2	0	1
Proline	0	1	1	2	0	1	1	1
Serine	0	1	63	1	2	1	1	0
Threonine	1	1	1	64	1	1	1	1
Tryptophan	0	1	1	1	0	0	1	0
Tyrosine	1	1	1	0	1	1	1	1
Valine	0	1	1	0	0	1	2	4

TABLE 4.5: Amino acid occurrences in training and testing nona-peptide data sets for CoEPrA Peptide Binding Affinity Task 3.

Training

Amino Acid	**Location**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	10	3	15	6	16	14	17	12	22
Arginine	5	0	1	8	3	4	3	1	0
Asparagine	2	0	4	6	3	4	3	0	0
Aspartic acid	1	0	10	9	5	3	0	5	0
Cysteine	2	1	2	1	1	2	2	4	1
Glutamine	1	0	1	13	2	4	4	1	0
Glutamic acid	0	0	2	4	4	3	3	6	0
Glycine	10	0	10	15	19	9	1	9	0
Histidine	1	0	2	2	5	1	2	4	0
Isoleucine	14	13	6	4	5	6	11	5	15
Leucine	17	88	22	10	15	16	16	29	33
Lysine	2	0	0	6	1	1	0	1	0
Methionine	5	10	7	1	2	6	2	3	0
Phenylalanine	16	0	7	4	10	6	19	11	0
Proline	1	0	4	20	5	26	8	5	0
Serine	13	0	9	9	1	5	7	16	0
Threonine	5	9	5	8	6	8	6	12	2
Tryptophan	4	0	8	3	4	2	1	2	0
Tyrosine	19	0	12	1	5	1	7	4	0
Valine	5	9	6	3	21	12	21	3	60

Testing

Amino Acid	**Location**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	17	6	17	8	17	6	16	19	27
Arginine	7	0	0	3	3	0	1	1	1
Asparagine	2	0	1	1	2	5	4	2	0
Aspartic acid	2	0	8	7	11	2	3	0	0
Cysteine	0	0	2	5	1	4	3	4	0
Glutamine	3	1	2	17	3	7	4	3	0
Glutamic acid	0	0	4	4	2	1	0	3	0
Glycine	10	0	4	23	21	8	3	9	0
Histidine	5	0	3	3	6	2	1	5	0
Isoleucine	16	4	6	1	4	5	4	6	14
Leucine	15	87	21	9	15	26	17	22	34
Lysine	4	0	2	5	1	1	3	1	0
Methionine	3	15	8	1	1	3	3	1	2
Phenylalanine	13	0	9	3	8	5	18	3	0
Proline	0	1	3	9	1	24	11	6	0
Serine	4	0	7	12	6	4	8	20	0
Threonine	1	7	4	6	4	11	8	13	2
Tryptophan	3	0	6	0	3	1	4	5	0
Tyrosine	16	0	18	3	4	5	3	2	0
Valine	12	12	8	13	20	13	19	8	53

TABLE 4.6: Amino acid occurrences in training and testing nona-peptide data sets for CoEPrA Peptide Binding Affinity Task 4.

Training

Amino Acid	Location								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	10	3	15	6	16	14	17	12	22
Arginine	5	0	1	8	3	4	3	1	0
Asparagine	2	0	4	6	3	4	3	0	0
Aspartic acid	1	0	10	9	5	3	0	5	0
Cysteine	2	1	2	1	1	2	2	4	1
Glutamine	1	0	1	13	2	4	4	1	0
Glutamic acid	0	0	2	4	4	3	3	6	0
Glycine	10	0	10	15	19	9	1	9	0
Histidine	1	0	2	2	5	1	2	4	0
Isoleucine	14	13	6	4	5	6	11	5	15
Leucine	17	88	22	10	15	16	16	29	33
Lysine	2	0	0	6	1	1	0	1	0
Methionine	5	10	7	1	2	6	2	3	0
Phenylalanine	16	0	7	4	10	6	19	11	0
Proline	1	0	4	20	5	26	8	5	0
Serine	13	0	9	9	1	5	7	16	0
Threonine	5	9	5	8	6	8	6	12	2
Tryptophan	4	0	8	3	4	2	1	2	0
Tyrosine	19	0	12	1	5	1	7	4	0
Valine	5	9	6	3	21	12	21	3	60

Testing

Amino Acid	Location								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	3	0	0	5	2	2	0	4	0
Arginine	1	1	1	2	3	0	1	0	0
Asparagine	1	0	9	1	4	4	0	2	0
Aspartic acid	1	0	0	4	0	0	4	2	0
Cysteine	1	2	0	0	0	0	5	0	3
Glutamine	1	1	0	1	2	12	2	0	0
Glutamic acid	1	0	0	8	4	1	1	0	1
Glycine	2	0	15	7	17	0	0	9	2
Histidine	0	0	0	0	0	0	0	2	0
Isoleucine	3	3	3	0	0	2	4	0	0
Leucine	0	31	5	2	3	6	0	4	9
Lysine	11	0	1	0	3	0	2	3	2
Methionine	0	5	0	0	0	3	2	0	0
Phenylalanine	4	0	3	2	3	0	2	4	0
Proline	0	0	0	8	0	1	8	1	0
Serine	1	0	0	0	4	0	2	8	1
Threonine	0	0	3	0	2	2	3	0	0
Tryptophan	0	0	2	2	0	0	3	0	0
Tyrosine	11	0	5	5	0	4	7	7	0
Valine	6	4	0	0	0	10	1	1	29

4.2.2 Mouse Class I MHC Peptide Binding Affinity Data Sets

Publicly available mouse class I MHC alleles (H2-Db, H2-Kb and H2-Kk) are used in this research study in order to find their real-value MHC-peptide binding affinities [286]. The allergenic regions of protein recognized by the binding site of any antibody are called epitopes (antigen derived peptides) [287], [288]. The epitopes in each allele contain experimentally measured binding affinities, numerically as pIC50. Each epitope in the data sets was represented by assigning values of physico-chemical or bio-chemical descriptors to each amino acid. The same set of descriptors (real values) for each amino acid aformentioned previously are used. As shown in Table 4.7, H2-Db consists of nona-peptides that have a total of 5787 (643x9 = 5787) descriptors, H2-Kb and H2-Kk have octa-peptides that have a total of 5144 (643x8 = 5144) descriptors. The statistics (range, mean and standard deviation) of the binding affinities of the mouse class I MHC alleles are given in Table 4.8.

TABLE 4.7: General characteristics of the data sets used for the prediction of peptide binding affinity for mouse class I MHC alleles.

Data Sets	Number of Peptide Sequences	Nature of Peptide	Number of Peptide Sequence Descriptors
$H2-D^b$	65	nona-peptide	5787
$H2-K^b$	62	octa-peptide	5144
$H2-K^k$	154	octa-peptide	5144

TABLE 4.8: The statistics of the binding affinity of mouse class I alleles.

Data Sets	Min	Max	Mean	Std
H2-Db	3.3570	8.6990	6.5428	1.2656
H2-Kb	3.8100	9.2220	6.8489	1.3441
H2-Kk	4.1920	8.4030	7.5231	0.8257

Table 4.9 - Table 4.11 shows the distribution of amino acids placed on the peptide locations for each of the mouse class I alleles. Data set analysis of these allele shows some strong preferences on various peptide locations. For the mouse class I H2-Db allele, Asparagine (N) at position 5 contributes very strongly on this allele with occupancy rate of 61. Serine (S) at position 1, Isoleucine (I) at positions 3 and 9, Glutamic acid (E) at positions 4 and 7, Leucine (L) at position 6 and 9, Methionine (M) at position 9 are

also strongly contribute to their positions with occupancy rate of more than 15. For the mouse class I H2-Kb allele, Leucine (L) at position 8 contributes very strongly on this allele with occupancy rate of 45. Phenylalanine (F) and Tyrosine (Y) are strongly contributing to position 5, with occupancy rates of 30 and 21, respectively. Serine (S) at position 1, Tyrosine (Y) and Isoleucine (I) at position 3 are also strongly contribute to their positions with occupancy rate of more than 10. At positions 1, 4, 6 and 7, amino acids are almost equally contributes to their positions with occupancy rate of less than 10. For the mouse class I H2-Kk allele, different amino acids very strongly dominate their positions with very high occupancy rates. Phenylalanine (P) at position 1, Glutomic acid (E) at position 2, Threonine (T) at position 3, Tryptophan (W) at position 4, Glycine (G) at position 5, Asparagine (N) at position 6, Leucine (L) at position 7, Isoleucine (I) at position 8 are contributing to their positions with occupancy rates of 130, 130, 128, 127, 130, 127, 130, 113, respectively.

TABLE 4.9: Amino acid occurrences for the H2-Db allele.

Amino Acid	**Location**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	7	11	3	3	1	6	4	8	2
Arginine	3	1	0	1	0	1	2	1	0
Asparagine	1	0	6	3	61	0	3	3	0
Aspartic acid	0	1	2	1	0	3	7	6	0
Cysteine	2	1	1	2	0	0	1	2	1
Glutamine	3	2	1	3	0	2	2	0	0
Glutamic acid	0	4	0	17	0	0	16	2	0
Glycine	1	6	4	4	3	5	4	2	0
Histidine	0	0	0	1	0	0	0	0	0
Isoleucine	6	3	15	3	0	3	3	5	15
Leucine	4	8	5	1	0	16	3	5	27
Lysine	2	3	2	2	0	2	1	1	0
Methionine	0	9	0	0	0	3	1	1	16
Phenylalanine	5	2	3	2	0	4	0	1	0
Proline	0	0	5	4	0	3	2	0	0
Serine	17	11	4	4	0	8	5	1	0
Threonine	4	3	0	6	0	1	2	10	0
Tryptophan	1	0	0	1	0	0	4	2	0
Tyrosine	5	0	3	3	0	2	2	14	1
Valine	4	0	11	4	0	6	3	1	3

TABLE 4.10: Amino acid occurrences for the H2-Kb allele.

Amino Acid	**Location**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	5	0	1	1	1	3	8	0
Arginine	6	0	1	4	0	4	7	0
Asparagine	3	3	0	9	0	1	6	0
Aspartic acid	1	3	0	1	0	4	0	0
Cysteine	1	0	0	1	0	4	0	0
Glutamine	2	2	4	5	0	7	3	0
Glutamic acid	0	1	2	3	0	3	0	0
Glycine	2	5	1	2	1	2	9	0
Histidine	3	1	2	2	1	0	2	0
Isoleucine	6	8	13	4	1	3	5	5
Leucine	8	4	4	7	2	8	6	45
Lysine	2	1	0	3	0	3	4	0
Methionine	6	2	0	1	0	0	0	5
Phenylalanine	3	1	4	4	30	1	1	0
Proline	0	4	1	2	0	7	3	0
Serine	7	14	6	5	1	6	4	0
Threonine	1	4	4	0	2	4	0	0
Tryptophan	0	1	0	3	1	0	1	0
Tyrosine	2	2	15	1	21	0	1	0
Valine	4	6	4	4	1	2	2	7

TABLE 4.11: Amino acid occurrences for the H2-Kk allele.

Amino Acid	**Location**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	2	4	1	1	2	1	1	1
Arginine	1	1	1	1	1	1	1	1
Asparagine	2	1	1	1	2	127	1	19
Aspartic acid	2	1	1	1	1	1	2	1
Cysteine	0	0	0	0	0	0	1	0
Glutamine	2	1	1	1	1	1	1	1
Glutamic acid	1	130	1	1	1	2	2	1
Glycine	1	2	1	2	130	3	1	1
Histidine	2	1	2	2	1	1	1	1
Isoleucine	1	1	3	1	1	1	2	113
Leucine	2	2	2	2	1	2	130	2
Lysine	1	1	2	2	1	3	1	1
Methionine	1	1	1	1	1	2	1	1
Phenylalanine	130	1	4	1	2	2	1	1
Proline	1	1	1	3	1	1	2	1
Serine	1	1	128	2	3	1	1	1
Threonine	1	2	1	127	1	1	1	1
Tryptophan	1	1	1	2	1	1	1	1
Tyrosine	1	1	1	1	1	1	1	1
Valine	1	1	1	2	2	2	2	5

4.2.3 Encoding Feature Space with Amino Acids based Features

An amino acid index is formed of twenty real-values that discriminates each amino acid in terms of their specificity and characteristics of a particular physio-chemical or bio-chemical property of a protein. The amino acid indices are derived from laboratory and computational experiments. AAindex Database is the well known database, with the latest update in 2008, that consists of 544 amino acid indices [289]. The collection of 544 amino acid indices are located on GenomeNet website and given with their reference information.

AA scales in this research study, however, are formed of 643 scales as these scales are obtained from the publicly available high-dimensional peptide datasets provided at the Comparative Evaluation of Prediction Algorithms modeling competition. So, the octa-peptides are encoded as 5144 (643x8) descriptors and the nona-peptides are encoded as (643x9) descriptors. The feature encoding process for octa-peptides and nona-peptides are illustrated in Fig. 4.1 and Fig. 4.2. Nevertheless, the data sets provided are lack of the definitions of these scales. The notes given with the data sets only tell that most of the indices are from AAindex Database and the remaining ones are from the literature. In order to reveal the definitions of these scales, the numeric values of each AA scale and its corresponding definition are searched from the AAindex Database and from the literature. It is discovered that most of them but not all of them are from the AAindex database. A total of 507 out of 643 amino acid indices are obtained from this database. However, many of the indices remain still unknown. The name and scales that are discovered after the searching process is broadly provided in Appendix A (Amino Acid Indices) and Appendix B (Amino Acid Scales).

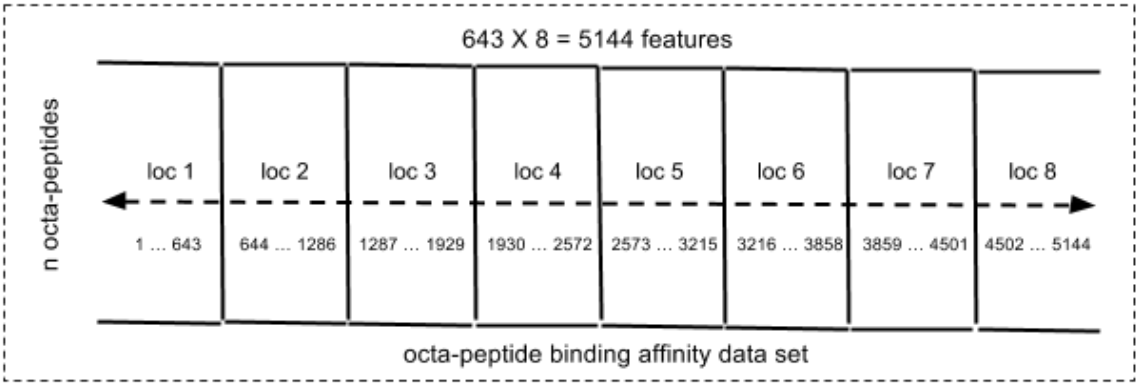


FIGURE 4.1: Feature encoding process for a octa-peptide.

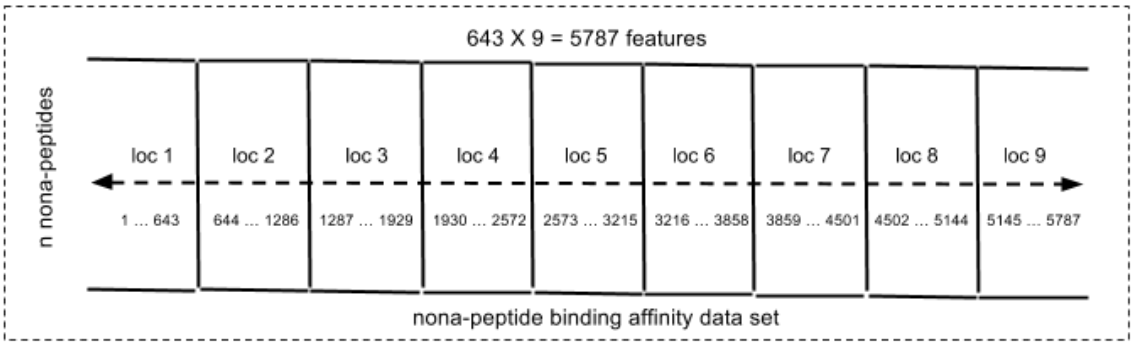


FIGURE 4.2: Feature encoding process for a nona-peptide.

The feature space for the peptide binding affinity data sets is encoded with 643 scales corresponding to each amino acid location on the peptide. At this step, 643 scales are transformed into their normalized values as shown in (4.1). The normalization helps to protect descriptors which have smaller variance value from those descriptors which have larger variance value as they may have more influence in the model building process. Additionally, all index values become standardized and proportional respect to each other. Unity-based normalization is used as the normalization method and the scales are normalized using a linear transformation. In the end, each scale normalized to a value in the interval [0, 1]. The unity-based normalization is computed as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{4.1}$$

where x, $max(x)$, $min(x)$ denote the index value, max value and min value for a typical amino acid index, respectively.

4.3 Results and Discussion

The feature selection is carried out by using the multi-cluster feature selection method to be able to derive the most significant feature subsets of the entire feature space containing around 5000 attributes. MCFS is an unsupervised feature selection method that does not require output or target label in the selection process [273]. Instead of a target label, it uses multiple eigenvectors of graph Laplacian. The number of used eigenvectors is set to the the number of features to be selected in this research study. MCFS requires the number of nearest neighbours parameter (k) for constructing the k-nearest-neighbours graph. The parameter value for the k is set to 5 (default). MCFS was able to deal with large number of attributes for the data sets efficiently. The reduced feature subset was used as input variables of the proposed rule-based fuzzy systems. The low dimensional structure is then expected to help eliminate noise in the data sets and provide more robust predictive models. However, the data sets can be exposed to the risk of information loss during the feature selection process.

In order to assess the importance of the features, the feature selection method was run separately to select the number of features between 1 and 250. More representative descriptors found in the 250 separate subsets seem to be repeatedly selected in each of the model's reduced features. The histograms for the selected features of each peptide binding affinity data set are shown in Fig. 4.3 - 4.8 presenting which molecular descriptors are strongly or weakly correlated with the binding affinity. Feature index represents the index of any descriptor that is encoded with 643 scales corresponding to each amino acid location on the peptide. Depending on the type of peptide, it is the position of an AA scale located between the first and last descriptor of the designated data set. The last descriptors are the 5144th and 5787th indexes for the octa-peptide and nona-peptide data sets, respectively. Number of occurrence shows that how many times a descriptor is selected among the 250 separate feature selection processes. For the CoEPrA peptide binding affinity data sets, feature selection is implemented on the training data sets. Task 3 and 4 use the same training data sets but they have separate data sets for the evaluation of their predictive models. The number of features that are appeared distinctly among the 250 feature selection steps are 398, 294, and 643 for Task1, Task 2 and Task 3-4, respectively. Corresponding to the indices of the descriptors that were selected highest were 2229 (AAindexID = 300) and 5524 (AAindexID = 380)

for the first task, 4294 (AAindexID = 436) for the second task, 4939 (AAindexID = 438) for the third and fourth tasks, respectively. Frequency of top ten selected amino acid indices are given in Table 4.12 - Table 4.14. For the mouse class I MHC alleles, feature selection is implemented on the entire data sets. The number of distinctly selected features among the 250 feature selection steps are 356, 370, and 424 for H2-Db, H2-Kb and H2-Kk, respectively. Corresponding to the indices of the descriptors that were selected highest were 1313 (AAindexID = 27) for the H2-Db, 1974 (AAindexID = 45) for the H2-Kb, 2365 (AAindexID = 436) for the H2-Kk, respectively. Frequency of top ten selected amino acid indices are given in Table 4.15 - Table 4.17. The descriptions of amino acids based features are provided in Appendix A.

Results show that the features selected for each data set are very different from each other. There is no common feature for the top ten most frequent features among data sets. One reason for this is that encoded features are mainly dependent on the peptides found in the data sets.

4.4 Conclusion

In this chapter, two groups of peptide binding affinity are studied. First group of data sets are the CoEPrA peptide binding affinity data sets that are formed of four tasks. Second group of data sets are the mouse class I MHC peptide binding affinity data sets (H2-Db, H2-Kb and H2-Kk). Amino acid occurrences of peptide data sets are provided in order to present the amino acid composition of each data set. To propose the predictive models, the feature space of the peptide data sets is encoded using the numerical values of bio-chemical descriptors corresponding to each amino acid location on the peptide. As each amino can be represented with a high number of descriptors, the encoded peptide data sets become high-dimensional data sets. In order to derive significant descriptors of these data sets, feature selection is applied. The low-dimensional representation of the proposed models allowed the elimination of noise and removal of redundant features. The selected features showed which molecular descriptors are strongly or weakly correlated with the binding affinity for the particular data set. Finally, it should also be noted that the features used to propose the predictive models in this thesis may not be the best representative feature sets. However, there might be better methods but current results seem to be very promising.

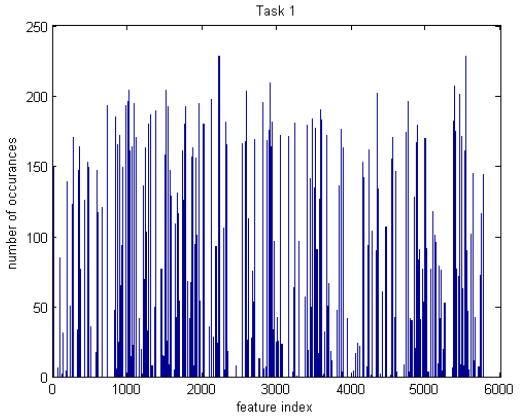


FIGURE 4.3: Number of occurrences of the selected features for Task 1.

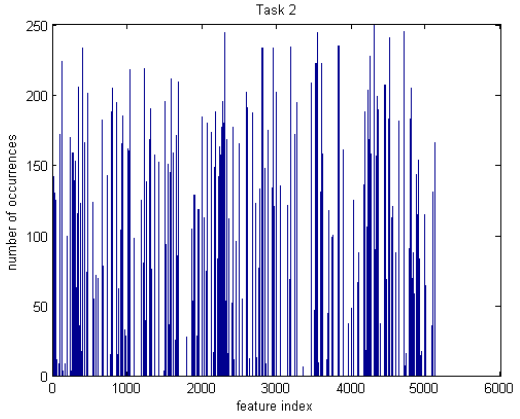


FIGURE 4.4: Number of occurrences of the selected features for Task 2.

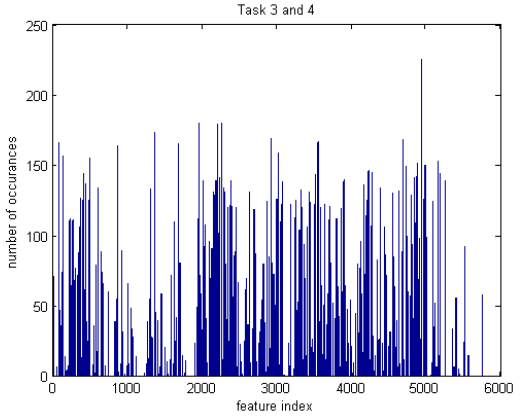


FIGURE 4.5: Number of occurrences of the selected features for Task 3 and 4.

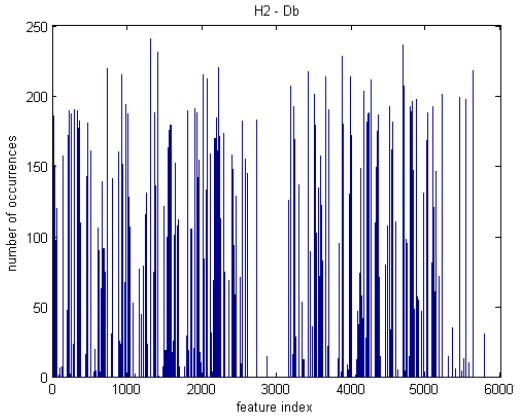


FIGURE 4.6: Number of occurrences of the selected peptide descriptors for H2-Db.

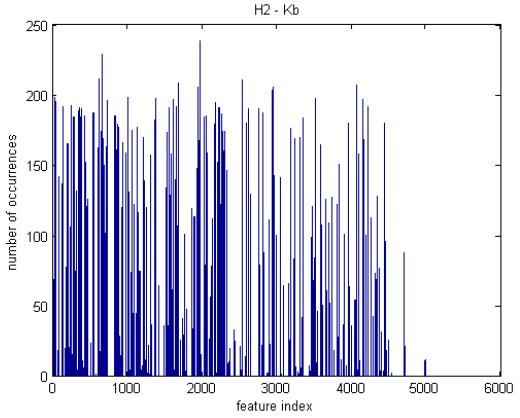


FIGURE 4.7: Number of occurrences of the selected peptide descriptors for H2-Kb.

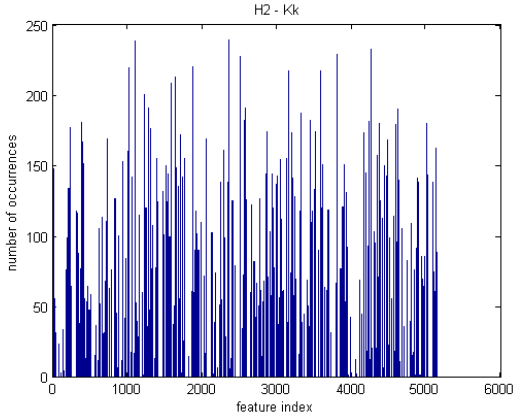


FIGURE 4.8: Number of occurrences of the selected peptide descriptors for H2-Kk.

TABLE 4.12: Top ten most frequent amino acid indices selected for Task 1.

No	Feature Index	Amino Acid Index		Number of Occurrences
		Index ID	Location	
1	2229	300	4	228
2	5524	380	9	228
3	2917	345	5	209
4	5379	235	9	207
5	1030	387	2	204
6	1515	229	3	204
7	2599	27	5	203
8	4339	481	7	202
9	5446	302	9	201
10	2125	196	4	197

TABLE 4.13: Top ten most frequent amino acid indices selected for Task 2.

No	Feature Index	Amino Acid Index		Number of Occurrences
		Index ID	Location	
1	4294	436	7	250
2	4697	196	8	245
3	2306	377	9	244
4	3539	324	6	244
5	4509	8	8	240
6	3826	611	6	235
7	3181	609	5	234
8	400	400	1	233
9	2807	235	5	233
10	2952	300	5	233

TABLE 4.14: Top ten most frequent amino acid indices selected for Task 3 - 4.

No	Feature Index	Amino Acid Index		Number of Occurrences
		Index ID	Location	
1	4939	438	8	225
2	1957	28	4	180
3	2267	338	4	180
4	2215	286	4	179
5	1374	88	3	173
6	2921	349	5	169
7	4689	188	8	168
8	3553	338	6	167
9	89	89	1	166
10	3550	335	6	166

TABLE 4.15: Frequency of amino acid indices that were selected highest for H2-Db.

No	Feature Index	Amino Acid Index		Number of Occurances
		Index ID	Location	
1	1313	27	3	240
2	4689	188	8	236
3	1406	120	3	231
4	3876	18	7	228
5	2224	295	4	220
6	731	88	2	219
7	5625	481	9	218
8	3420	205	6	217
9	924	281	2	215
10	2017	88	4	215

TABLE 4.16: Frequency of amino acid indices that were selected highest for H2-Kb.

No	Feature Index	Amino Acid Index		Number of Occurances
		Index ID	Location	
1	1974	45	4	238
2	661	18	2	229
3	628	628	1	211
4	2538	609	4	210
5	1686	400	3	208
6	4066	208	7	207
7	1947	18	4	205
8	2952	380	5	205
9	2939	367	5	203
10	2936	364	5	201

TABLE 4.17: Frequency of amino acid indices that were selected highest for H2-Kk.

No	Feature Index	Amino Acid Index		Number of Occurances
		Index ID	Location	
1	2365	436	4	239
2	1105	462	2	238
3	4258	400	7	232
4	3801	586	6	229
5	2515	586	4	227
6	1872	586	3	220
7	1019	376	2	219
8	3158	586	5	217
9	3579	364	6	217
10	1650	364	3	213

Chapter 5

Quantitative Prediction of Peptide Binding Affinity with SVR-based Type-1 Fuzzy System

5.1 Introduction

Peptide binding plays vital roles in many biological processes such as activating the cytotoxic T-cells in the immune system. The T-cell receptor is a molecule, present at the T-cell surface, and signicantly required to activate the T-cell by recognising antigenic peptides bound to MHC molecules translocated on the surface of the infected cells. The peptide epitopes that are bound to MHC class I molecules can be recognised by the T-cells and can induce the cellular immune response.

Support vector regression is one of the earliest quantitative approaches that is proposed to model MHC-peptide complex for finding precise binding affinities [25]. This approach as a non-linear method has achieved a better performance compared to linear models such as the additive method [290]. The non-linear modeling approach has been taken by a number of later methods such as regularization methods [81], partial least squares [291] and random forests [292] to reveal the real-value of the binding affinity. SVM is one of the computational methods that has been shown to effectively deal with large number of dimensions [157]. When the quantitative modelling is the case, SVMs can be extended to SVR with the aid of e-sensitive loss function [221]. SVR has been proven to

lead better generalization ability and performance in a wide range of applications [25], [293]. Fuzzy systems is another non-linear method that is good at modelling uncertainty and yielding a set of interpretable if-then rules [168]. On the contrary, fuzzy systems can suffer from the curse of dimensionality in high-dimensional systems.

Roughly speaking, general frameworks that incorporates fuzzy systems with the support-vector based methods fall into two approaches. The first approach is to extract support vectors from the training data set to generate fuzzy rule-base [294], [295], [296]. The second approach is to employ support vector mechanism to learn consequent parameters of the fuzzy system [297]. Recent efforts for the second approach focused for the design of a general framework similar to the layered structure of neuro fuzzy systems [298], [299], [300]. In this chapter a hybrid computational model support-vector based TSK fuzzy system (TSK-SVR I) that follows the second approach, is presented and applied to effectively model quantitative prediction of binding affinities between major histo-compatibility complex proteins and peptides which is an important problem in biology and medicine with applications for drug design.

In the next section, a proposed type-1 fuzzy system is described in detail. In Section 5.3 the results of the binding affinity problem are presented and discussed. Finally, Section 5.4 draws the conclusions of this chapter.

5.2 Materials and Methods

In this section, a proposed type-1 fuzzy system is described over the following subsections: TSK Type-1 Fuzzy System (5.2.1), Generating Fuzzy System with Fuzzy Clustering (5.2.2), SVR-based Type-1 TSK Fuzzy System (5.2.3).

5.2.1 Type-1 TSK Fuzzy System

Each rule in the structure of the TSK fuzzy system can be expressed in the following form [6]:

$$\begin{aligned} &R_i : \text{IF } x_1 \text{ is } A_{1i} \text{ AND } x_2 \text{ is } A_{2i} \text{ ... AND } x_n \text{ is } A_{ni} \\ &\text{THEN } y_i \;=\; a_{0i} + a_{1i}x_1 + ... + a_{ni}x_n \end{aligned} \tag{5.1}$$

where $i = 1..r$ is the number of fuzzy rules; and (x_1, x_2, ..., x_n) are the n input variables; and a fuzzy set for the variable n and rule i is denoted by A_{ni}; and y_i is the rule output of the consequent part; and a_{ni} represents the coefficient of its linear equation.

The fuzzy set A_{ij} is described with any form of membership functions, commonly with the following Gaussian membership function:

$$\mu(x_j) = e^{-\frac{(x_j - c_{ij})^2}{2(\sigma_{ij})^2}} \tag{5.2}$$

where $\mu(x_j)$ is the degree of membership for input variable x_j; and c_{ij} and σ_{ij} are the centre and standard deviation that characterises a fuzzy set, respectively. The t-norm operation can be defined as:

$$f_i = \prod_{j=1}^{n} \mu(x_j) \tag{5.3}$$

where f_i is the firing strength determined by using a t-norm operation defined by the product (*) operator. A normalised firing strength can be defined in the following form:

$$\overline{f_i} = f_i / \sum_{k=1}^{r} f_k \tag{5.4}$$

where $\overline{f_i}$ denotes normalised firing strength. A defuzzification operation is processed by finding the overall output obtained by the weighted sum:

$$y = \sum_{i=1}^{r} \overline{f_i} y_i \tag{5.5}$$

5.2.2 Generating Fuzzy System with Fuzzy Clustering

Fuzzy clusters are more flexible than the crisp clusters. In fuzzy clustering, each data sample in the data set is assigned a degree of membership for each of the partitions. Therefore, the memberships along with the mean values of each cluster obtained at the end of fuzzy clustering process can be used to derive the premise part of the fuzzy system. Thus, the outputs of the fuzzy clustering process can be used to approximate the membership functions that characterize each fuzzy set found in the rule-base and to identify structure of the fuzzy model [9], [10].

Fuzzy c-Means method partitions data set into a number of clusters in a way that each data object is assigned a degree of membership for each cluster [242]. The FCM model aims to minimise the optimisation function:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{\tau}\|x_j - c_i\|^2 \tag{5.6}$$

where $\tau \in (1,\infty)$ is the degree of fuzzification; n is the number of samples; c is the number of clusters, $2 \leq c \leq n$; $V = \{c_1, ..., c_n\}$ is the set of cluster prototypes; $c_i \in \mathrm{R^p}$ is the $\mathrm{i^{th}}$ point prototype; $\mathrm{u_{ij}}$ is the degree of membership of the $\mathrm{j^{th}}$ sample for the $\mathrm{i^{th}}$ point prototype; $U = [u_{ij}]$ is a c $\times$ n membership matrix.

The sum of membership values of an object is constrained to one. The clustering process iteratively calculates cluster centres and degrees of memberships of each data point until the J_m is satisfied or the number of iterations reaches a preset value:

$$\mathrm{c_i} = \frac{\sum_{j=1}^{n} u_{ij}^{\tau}\mathrm{x_j}}{\sum_{j=1}^{n} u_{ij}^{\tau}} \tag{5.7}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{||\mathrm{x_j}-\mathrm{c_i}||}{||\mathrm{x_j}-\mathrm{c_k}||}\right)^{\frac{2}{\tau-1}}} \tag{5.8}$$

5.2.3 SVR-based Type-1 TSK Fuzzy System

Fuzzy systems are able to model uncertain and imprecise knowledge and forms a structure for representing human reasoning. Usually, fuzzy systems can be constructed by obtaining knowledge from human experts. Nonetheless human experts may not be available all the time, and building a model using a classical non-linear system with a limited prior knowledge is often difficult [5]. Among the various fuzzy systems, TSK is commonly used for modelling complex systems [6], [7]. TSK is a fuzzy modelling method,

proposed by Takagi, Sugeno and Kang, that can exhibit high-dimensions, non-linearity, and complexity. TSK-FS can be combined with other methods, particularly learning methods, and enhanced with learning and adaptation capabilities [8].

In TSK models, rule antecedent is in the form of membership functions and the rule consequent is a linear function of inputs. Although there are many methods proposed to model TSK-FS, general approach is to keep the premise parameters constant whereas values of the consequent parameters are computed by the least square estimation which is a statistical modeling that assumes a linear relationship that exists between input and output variables. The performance of these models are often determined by how accurately the actual output value can be predicted from the input variables. This learning approach is based on minimising the empirical risk and constitutes an essential part of the fuzzy systems [301], [209]. One drawback of least squares learning algorithm is that even though the training error is minimised, the model can badly suffer from the overfitting. There are methods that have been explored for addressing the problems in the least square estimation. One of the methods is support vector regression [220], [221] that has been shown to be an efficient and robust method and provides high generalizability and performance. Applications of SVR have demonstrated considerably better modeling in various non-linear systems and minimising the structural risk than least squares approach. This concept can be incorporated with TSK-FS to better train the consequent part of the TSK-FS.

Let the input and real-valued output training data set D is $\{(\vec{x_1}, y_1),\ (\vec{x_2}, y_2),\ ...,\ (\vec{x_n}, y_n)\}$. In order to obtain the coefficients w (weight vector) and b (bias term) of the SVR linear expression, each data item $\vec{x_i}$ in the training data set along with its actual output y_i is transformed to represent a training data pair $(\vec{x_i}', y_i)$ which is fed into SVR as in the following form:

$$([\overline{f_i}, \overline{f_i}x_{i1}, \overline{f_i}x_{i2}, ..., \overline{f_i}x_{in}], y_i) \tag{5.9}$$

Once the w and b are obtained, a defuzzification operation for the support vector-based Takagi-Sugeno-Kang fuzzy system is formulated as:

$$y_i' = w_{0r} + \sum_{i=1}^{n}(w_{ir}x_i) \tag{5.10}$$

$$y' = \sum_{i=1}^{r} (\overline{f_i} y'_i + \frac{b}{r}) \tag{5.11}$$

where the new defuzzified output formulation of the SVR based type-1 TSK fuzzy model is denoted by y'. SVR part of the hybrid method is implemented through the use of LIBSVM package [302].

5.2.4 Predictive Modelling of Peptide Binding Affinity

This section presents the construction of SVR based type-1 TSK fuzzy models and identification of their parameters in the following steps. The SVR based type-1 TSK fuzzy model (TSK-SVR I) proposed for the prediction of peptide binding affinity is presented in Fig. 5.1.

5.2.4.1 Preprocessing

The model definition for the peptide binding affinity data sets started with turning amino acids of the peptides into numerical descriptors using amino acid indices. Then these numerical descriptors that form the data set is normalized in order for every feature to fall within the same range of values.

5.2.4.2 Feature Selection

To ease the processing of high-dimensionality of the input space of the fuzzy system, the number of features to be selected should be determined. The feature selection is carried out by using the Multi-Cluster Feature Selection method [273] to be able to derive the most significant feature subsets of the entire feature space containing around 5000 attributes. It should be noted that MCFS method itself suffers from the curse of dimensionality. Therefore, the number of features to be selected should be set as low as possible. In order to assess importance of the features, the feature selection method was run by using the predictive models separately between 1 and 250 features.

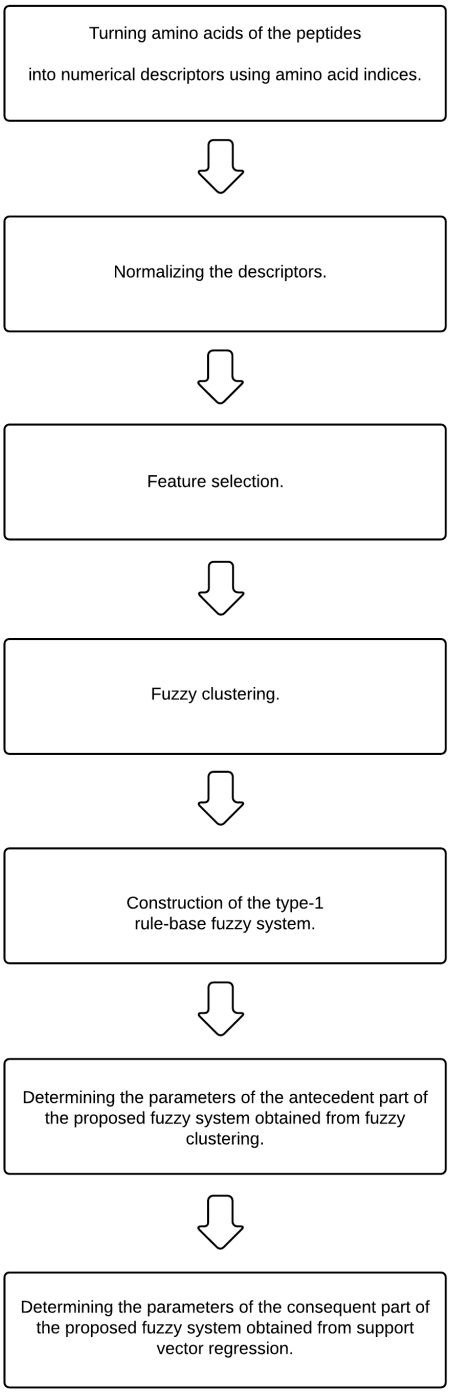


FIGURE 5.1: Stages of the SVR based type-1 TSK fuzzy model for the prediction of peptide binding affinity.

5.2.4.3 Identifying Antecedent Parameters

Fuzzy clustering is used as a pre-processing step to determine the antecedent parameters in fuzzy models. The parameter indicating the number of clusters should be preset before the fuzzy clustering is performed. The degree of fuzzification in fuzzy clustering is mostly chosen to be a value between 1.5 and 3 and set to two ($\tau = 2$) in this research study [242]. The number of clusters parameter is also used for determining the number of rules for the fuzzy system. The membership values and cluster prototypes obtained from fuzzy clustering is used to approximate the membership functions. The fuzzy sets involved in the rules are fully characterised by their membership functions. The parameters of membership functions obtained from these fuzzy clusters form the fuzzy sets of the premise part of TSK-FS.

5.2.4.4 Identifying Consequent Parameters

The rule consequent of TSK-FS is formed of linear function of inputs. Mainly, least squares method is used for finding the coefficients of linear functions. The least squares method is considered to be replaced by the support vector regression in this research study as it is more efficient and provides high generalizability and performance. For the consequent part of the fuzzy system, two parameters C and ϵ are required to be optimised for the SVR linear kernel.

5.2.4.5 Searching for Optimal Parameters

The number of clusters (parameter for the fuzzy clustering), ranging from two to seven is preset separately for each of the fuzzy clustering processes. The number of clusters determines the number of rules for the fuzzy model. The number of features to be selected is another parameter required to be set before processing the fuzzy model. For the consequent part of the fuzzy model as SVR is being used, two parameters (C and ϵ) are required to be set. In order to avoid the problem of overfitting, the parameters need to be selected properly. Due to the fact no generally accepted methods exist to determine these parameters optimally, the grid-search method has been decided to be employed as a parameter selection method in order to find the optimal parameter set. The grid-search method is simple and reliable and allows to implement parallel

computations. The parameters (C and ϵ) are searched within the given range with a step size of 0.05 to find out the optimal linear coefficients of the proposed model. For the features, the search range was decided to be between 1 and 250. It is hoped that these ranges broadly cover all the possibilities that may contain optimal measure. Therefore, these parameters as well as different combinations of the features are assessed and their results were presented. Fig. 5.2 depicts how the grid-search conducted on SVR kernel parameters (C and ϵ) for Tasks 1 - 4 for their given ranges and determined clusters and descriptors. Tables 5.1 - 5.3 show the optimal TSK-SVR I model parameter values of the proposed models for the peptide binding affinity data sets.

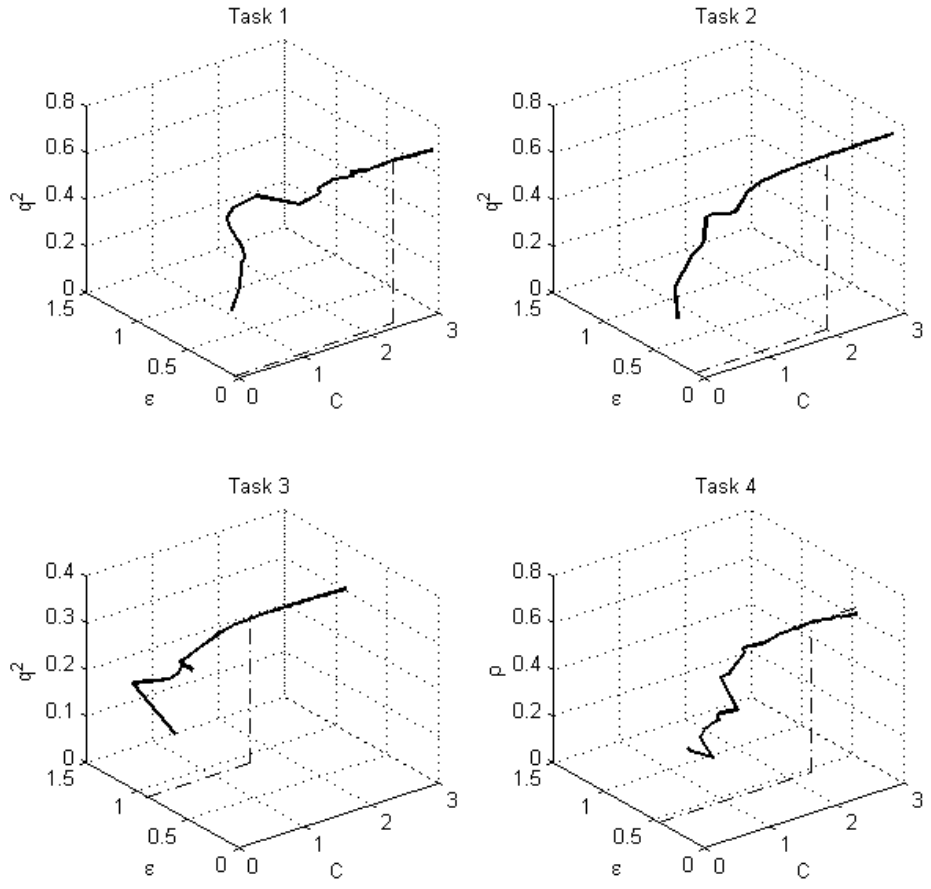


FIGURE 5.2: An example for the grid-search carried out to obtain the optimum values of linear SVR kernel parameters (C and ϵ) for peptide binding affinity Tasks 1-4.

TABLE 5.1: The optimal TSK-SVR I model parameter values for each peptide binding affinity data sets.

Task 1

number of clusters	number of selected features	C	ϵ
2	161	0.65	0.05
3	161	1.00	0.05
4	161	1.30	0.05
5	161	1.65	0.05
6	161	2.00	0.05
7	161	2.40	0.05

Task 2

number of clusters	number of selected features	C	ϵ
2	246	1.4	0.1
3	247	1.9	0.1
4	247	2.5	0.1
5	247	3.2	0.1
6	247	3.0	0.1
7	247	3.0	0.1

Task 3

number of clusters	number of selected features	C	ϵ
2	165	0.75	0.85
3	172	1.45	0.90
4	165	1.45	0.85
5	165	1.80	0.85
6	165	2.50	0.85
7	165	2.50	0.85

Task 4

number of clusters	number of selected features	C	ϵ
2	141	2.30	0.45
3	141	3.00	0.45
4	141	4.60	0.45
5	141	4.65	0.45
6	141	4.75	0.45
7	121	0.05	0.05

TABLE 5.2: The optimal TSK-SVR I model parameter values for each mouse class I allele entire data set prediction.

q^2

Allele	number of selected features	C	ϵ
H2-Db	30	75.0	0.20
H2-Kb	25	25.0	0.50
H2-Kk	62	18.5	0.20

AR

Allele	number of selected features	C	ϵ
H2-Db	39	9.75	0.05
H2-Kb	24	9.65	0.05
H2-Kk	22	7.50	0.05

TABLE 5.3: The optimal (q^2) TSK-SVR I model parameter values for each mouse class I allele leave-one-out cross validated prediction.

q^2			
Allele	number of selected features	C	ϵ
H2-Db	34	0.45	0.05
H2-Kb	32	0.25	0.15
H2-Kk	21	3.10	0.05

5.3 Results and Discussion

A non-linear system is proposed with the aid of support vector-based regression to improve the fuzzy system and applied to the real value prediction of degree of peptide binding. The experimental results and findings of the proposed method are validated using peptide binding affinity data sets that are different and independent from each other. Two groups of data sets are used for the performance evaluation and verification of the proposed approach that models the relationship between the peptides and their binding affinities. The first group of data sets consists of CoEPrA data sets. These data sets are used for the evaluation of the performance of the proposed method through blind-validation. The second group of data sets consists of mouse class I MHC alleles. These data sets are used for the evaluation of the performance of our method through cross-validation. The proposed model applied for each group separately. Compared to the previously published results in the literature, the proposed models yield an improvement in the prediction accuracy.

5.3.1 Blind-Validated Peptide Binding Affinity Prediction

There are some important parameters required to be set in antecedent and consequent parts that are likely to effect the performance of the fuzzy models. The parameters C and ϵ are used to optimise the SVR linear kernel for the consequent part. As previously mentioned, the proposed model (TSK-SVR I) was applied to four tasks and their optimal values of TSK-SVR I parameters (C and ϵ) were found using grid-search. The grid-search is repeated for each of the feature selection process (between 1 and 250 features). After, the each feature selection step, the best model for that step is selected. This process is repeated for different number rules (Fig. 5.3 - Fig. 5.8). The graphs show their corresponding prediction performances in terms of q^2 for the first three tasks and ρ for the last task. Solid line on graphs shows the separation of positive from negative q^2 values. Dashed line on graphs shows the highest q^2 value reached during the feature selection process. It should be noted that the cluster centers and the membership matrix is randomly initialized in the fuzzy clustering stage. Thereby, random initialization in FCM may have some effect on the performance. For Task 1, graph shows fluctuations and reaches three local maximums particularly in the first 100 features. It rose gradually then and reaches the global maximum at 161 features. After reaching the global maximum

it becomes steady. For Task 2, graph increases gradually as the number of features selected grew. It reaches two local maximums in the first 75 features and reaches the global maximum at around 247 features. For Task 3, slight fluctuations are observed through out the graph, reaching four local maximums in the first 150 features and then reaching global maximum at 172 features. For Task 4, substantial fluctuations are observed through out the graph, reaching three local maximums after 50 features until reaching global maximum at 141 features.

For each rule-base (rules that range between two and seven), feature selection (between 1 and 250 features) was carried out to reduce the number of features. It should be noted that selected features are highly dependent on their data sets. Approximately 5% of the features were sufficient for finding the optimal results. The amino acid features that contributed most to the efficiency of the proposed models are given in Table 5.4 - Table 5.7. For Task 1, eight amino acid features contributed to the output in more than four separate locations. The amino acid feature numbered with 481 (Hydrophobicity coefficient in reversed phase high performance liquid chromatography) contributed highest as it is represented in seven separate locations on each of the nona-peptide within the data set. This finding suggests that hydrophobic effect is important in mediating the binding process between the peptide and MHC molecule in this data set. Therefore, peptides can be shielded from the surrounding solvent and can be buried inner side of the protein [303]. For Task 2, eleven amino acid features contributed to the output in more than five separate locations. The amino acid feature numbered with 364 (Zimm-Bragg parameter sigma x 1.0E4) contributed highest as it is represented in seven separate locations on each of the octa-peptide within the data set. This finding suggests that helix formation in peptides is important in mediating the binding process between the peptide and MHC molecule in this data set. One main reason for the peptides that can nucleate a helix formation is that the ability of their side chains to participate in hydrophobic bonding [304]. For Task 3, nineteen amino acid features contributed to the output in more than three separate locations. The amino acid features numbered with 110 (Composition), 338 (Relative preference value at C"), 376 (Relative population of conformational state A), 405 (Normalized positional residue frequency at helix termini N") contributed highest as they are represented in four separate locations on each of the nona-peptide within the data set. For Task 4, ten amino acid features contributed to the output in more than three separate locations. The amino acid features numbered

with 306 (Average relative fractional occurrence in A0(i-1)), 338 (Relative preference value at C"), 110 (Composition), 125 (Normalized relative frequency of double bend) contributed highest as they are represented in seven separate locations on each of the nona-peptide within the data set. The amino acid feature numbered with 400 (Polarity) appeared in Task 1, Task 2 and Task 3 as a common feature with location occurrences of 4, 6 and 3, respectively. Therefore, the polarity of an amino acid is considered as one of the highly discriminating feature in these data sets. This finding suggests that polarity is important in mediating the binding process between the peptide and MHC molecule in this data set. It is reported that polarity of amino acids can play important role for the protein ubiquitination process. [305]. The full descriptions of amino acid features can be found in Appendix A.

Table 5.8 depicts prediction results based on the size of rule-base. Better results can also be achieved even with the reduced number of descriptors. The former value indicates the best prediction results obtained under the possible decreased feature set and the latter value shows the best performance at designated feature set. As the number of rules increased the results are improved for Task 1. For the remaining tasks there is no direct correlation is observed between the rule size and performance improvement. The experiments were also conducted with SVR alone. The optimal parameters depicted in Table 5.1 are also set for the SVR models. The SVR with a reduced feature subset yielded poorer results as compared the proposed method however outperformed the other SVR based methods in the literature as shown in Table 5.9. The predictive performance for Tasks 2, 3 and 4 have been improved by 15.9%, 28.8%, and 1.7%, respectively. For Task 1, no improvement gain is obtained.

For each rule-base the proposed method is able to build a robust and interpretable fuzzy system for a high-dimensional data set with a relatively small number of data samples. Table 5.10 depicts best prediction results as compared to the literature. For each task the results obtained are comparatively better than the recent studies presented in [81], [285], [291] and [292]. The predictive performance for Tasks 1, 2, 3 and 4 have been improved by 0.7%, 11.2%, 33.6% and 9.7% to the best model (depicted with boldface) presented in the literature, respectively. The overall improvement gain for all tasks is found to be 13.6%. The results also outperform the competition results in which each participant competed with their best model. In this competition Task 1 and 2 contained more than ten participants. Task 3 and 4 contained more than five participants.

The outcomes of the experiments clearly highlighted the strengths of TSK-SVR I. TSK-FS is more capable of managing uncertainty that exists in the data sets [5]. SVR based TSK-FS dealt with the curse of dimensionality effectively and yielded a better generalization performance [296], [300]. The results clearly suggest that the fuzziness has positively contributed towards the modeling of the tasks. The results also appear to suggest that different sets of variables effect the result, and that exploration of the feature selection methods may further help accelerate the predictive power of the proposed hybrid method.

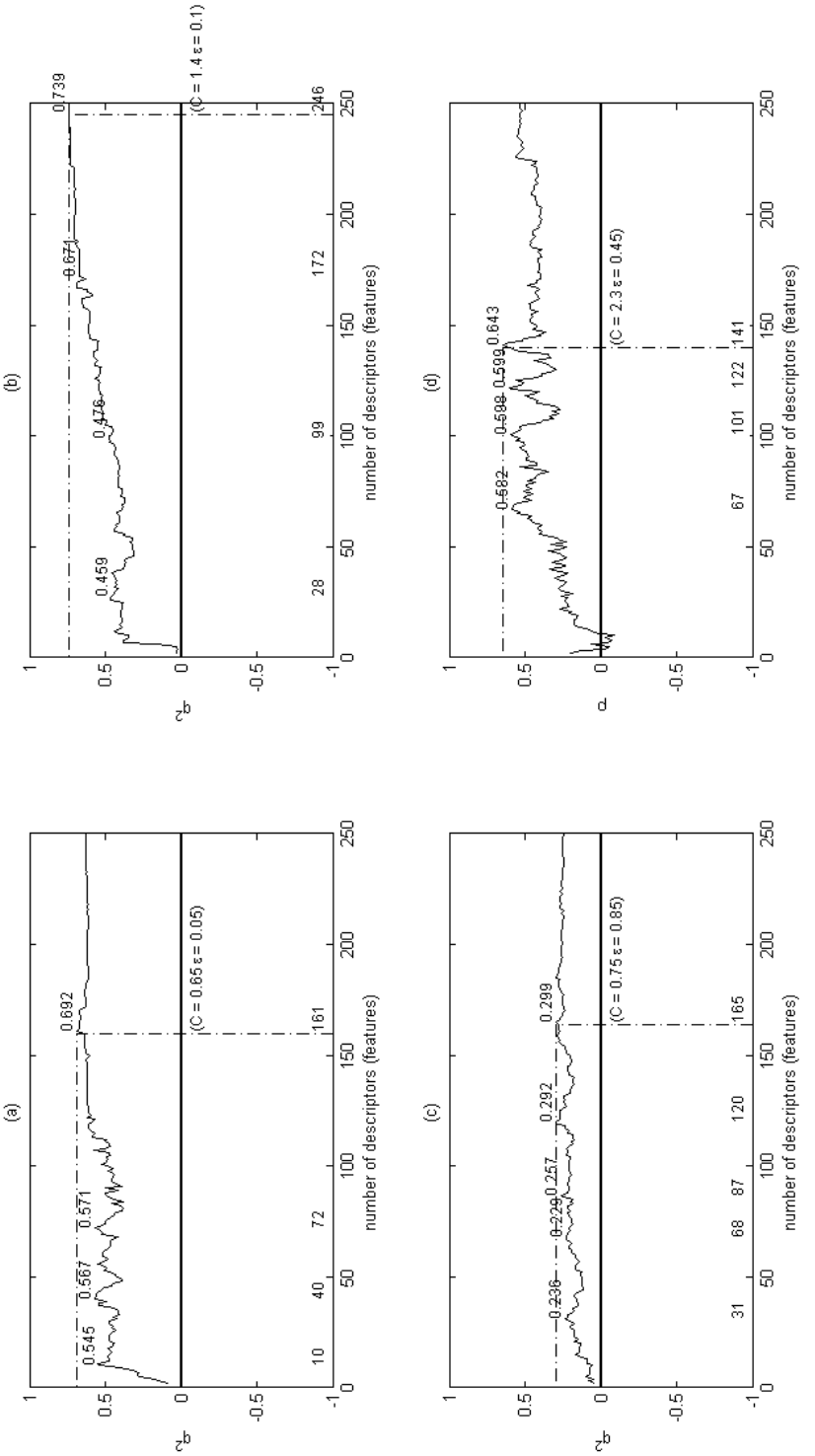


FIGURE 5.3: The performance of 2-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 0.65 and ϵ = 0.05). b) Task 2: Graph shows distinct peaks when the number of descriptors are 28, 99, 172 and reaches highest peak at 246 with the SVR parameters (C = 1.4 and ϵ = 0.1). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120 and reaches highest peak at 165 with the SVR parameters (C = 0.75 and ϵ = 0.85). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101, 122 and reaches highest peak at 141 with the SVR parameters (C = 2.3 and ϵ = 0.45).

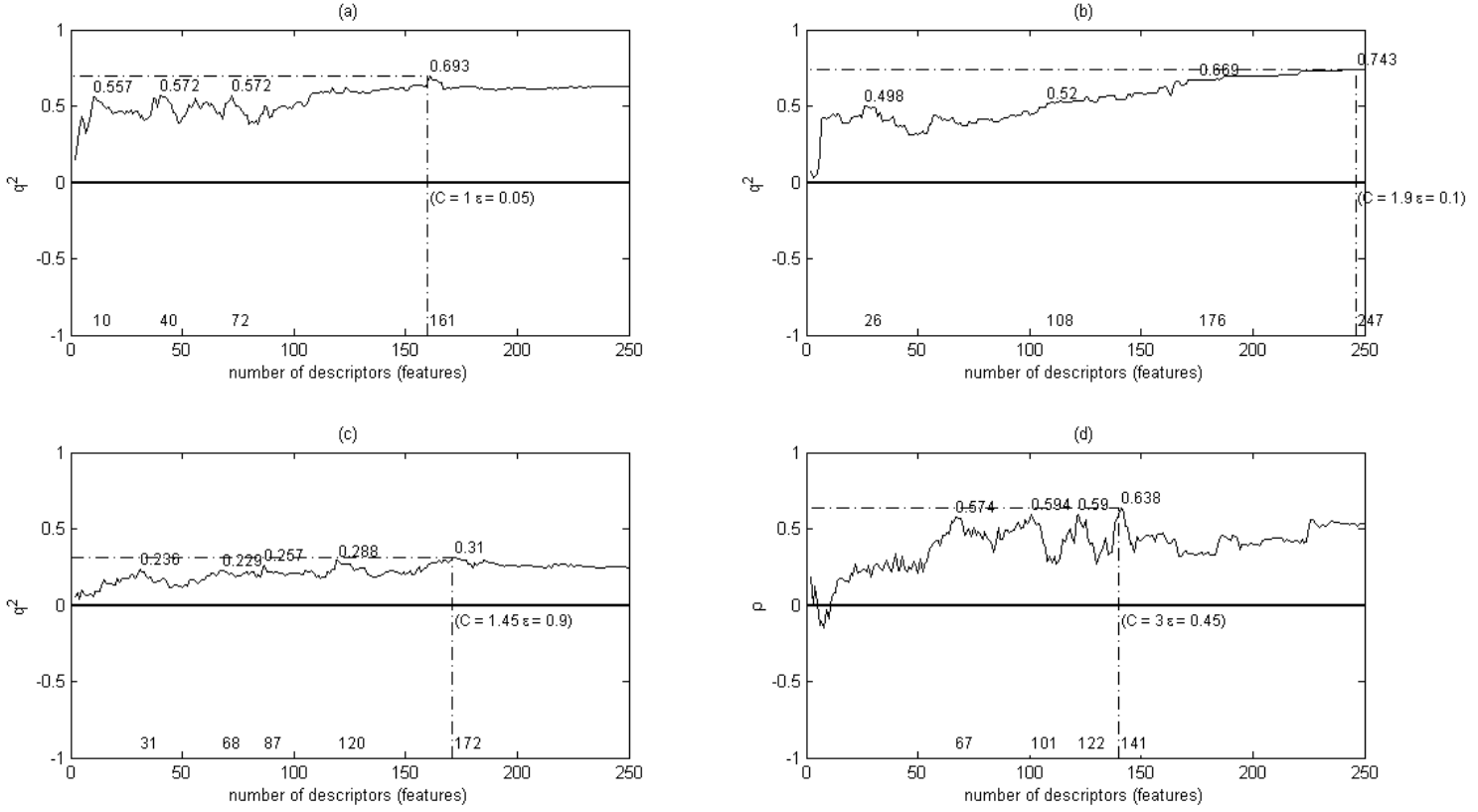


FIGURE 5.4: The performance of 3-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 1.0 and $\epsilon = 0.05$). b) Task 2: Graph shows distinct peaks when the number of descriptors are 26, 108, 176 and reaches highest peak at 247 with the SVR parameters (C = 1.9 and $\epsilon = 0.1$). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120 and reaches highest peak at 172 with the SVR parameters (C = 1.45 and $\epsilon = 0.9$). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101, 122 and reaches highest peak at 141 with the SVR parameters (C = 3.0 and $\epsilon = 0.45$).

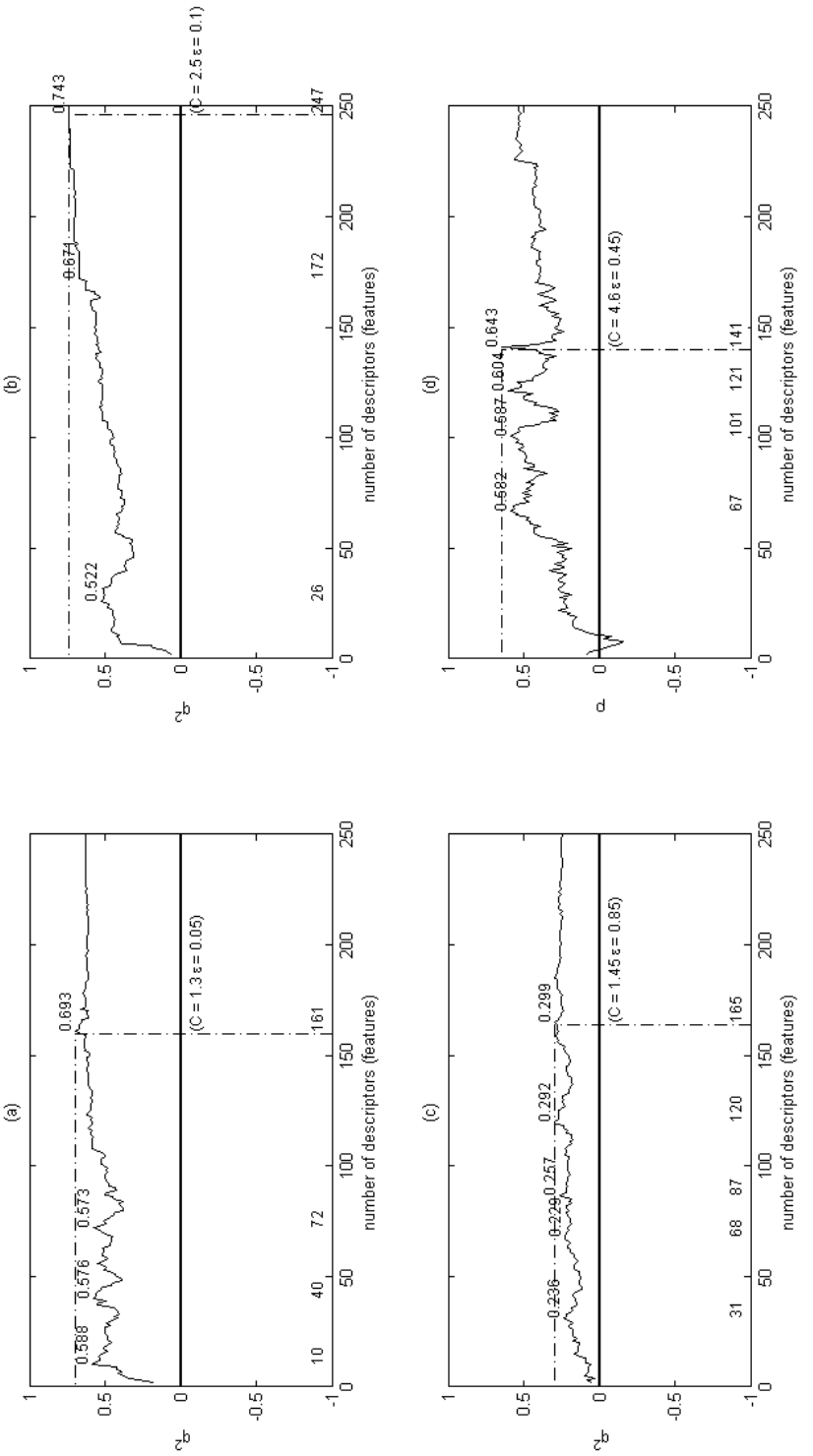


FIGURE 5.5: The performance of 4-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 1.3 and ϵ = 0.05). b) Task 2: Graph shows distinct peaks when the number of descriptors are 26, 172 and reaches highest peak at 247 with the SVR parameters (C = 2.5 and ϵ = 0.1). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120 and reaches highest peak at 165 with the SVR parameters (C = 1.45 and ϵ = 0.85). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101, 121 and reaches highest peak at 141 with the SVR parameters (C = 4.6 and ϵ = 0.45).

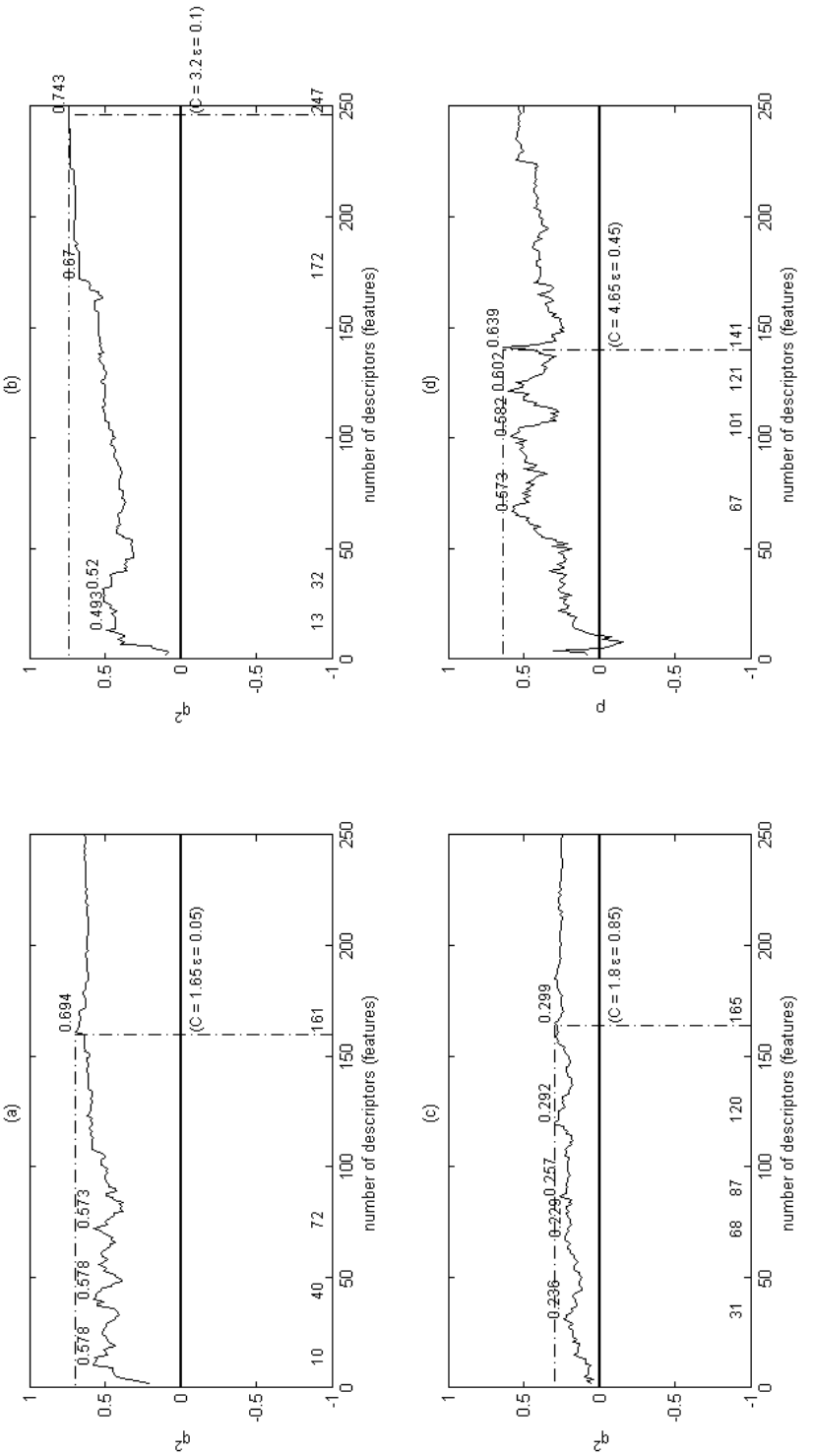


FIGURE 5.6: The performance of 5-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 1.65 and ϵ = 0.05). b) Task 2: Graph shows distinct peaks when the number of descriptors are 13, 32, 172 and reaches highest peak at 247 with the SVR parameters (C = 3.2 and ϵ = 0.1). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120 and reaches highest peak at 165 with the SVR parameters (C = 1.8 and ϵ = 0.85). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101, 121 and reaches highest peak at 141 with the SVR parameters (C = 4.65 and ϵ = 0.45).

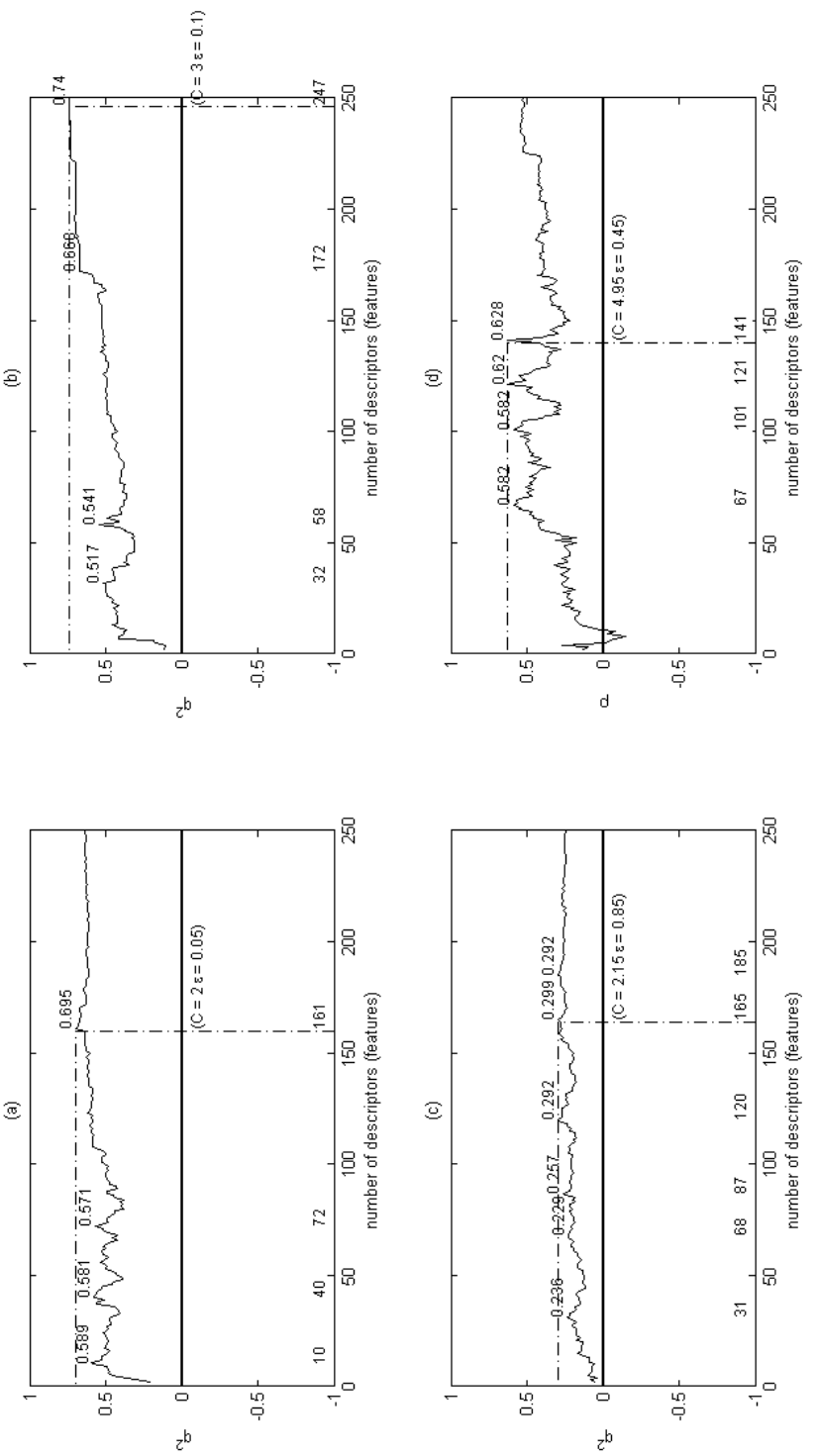


FIGURE 5.7: The performance of 6-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 2.0 and $\epsilon = 0.05$). b) Task 2: Graph shows distinct peaks when the number of descriptors are 32, 58, 172 and reaches highest peak at 247 with the SVR parameters (C = 3.0 and $\epsilon = 0.1$). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120 and reaches highest peak at 165 with the SVR parameters (C = 2.15 and $\epsilon = 0.85$). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101, 121 and reaches highest peak at 141 with the SVR parameters (C = 4.95 and $\epsilon = 0.45$).

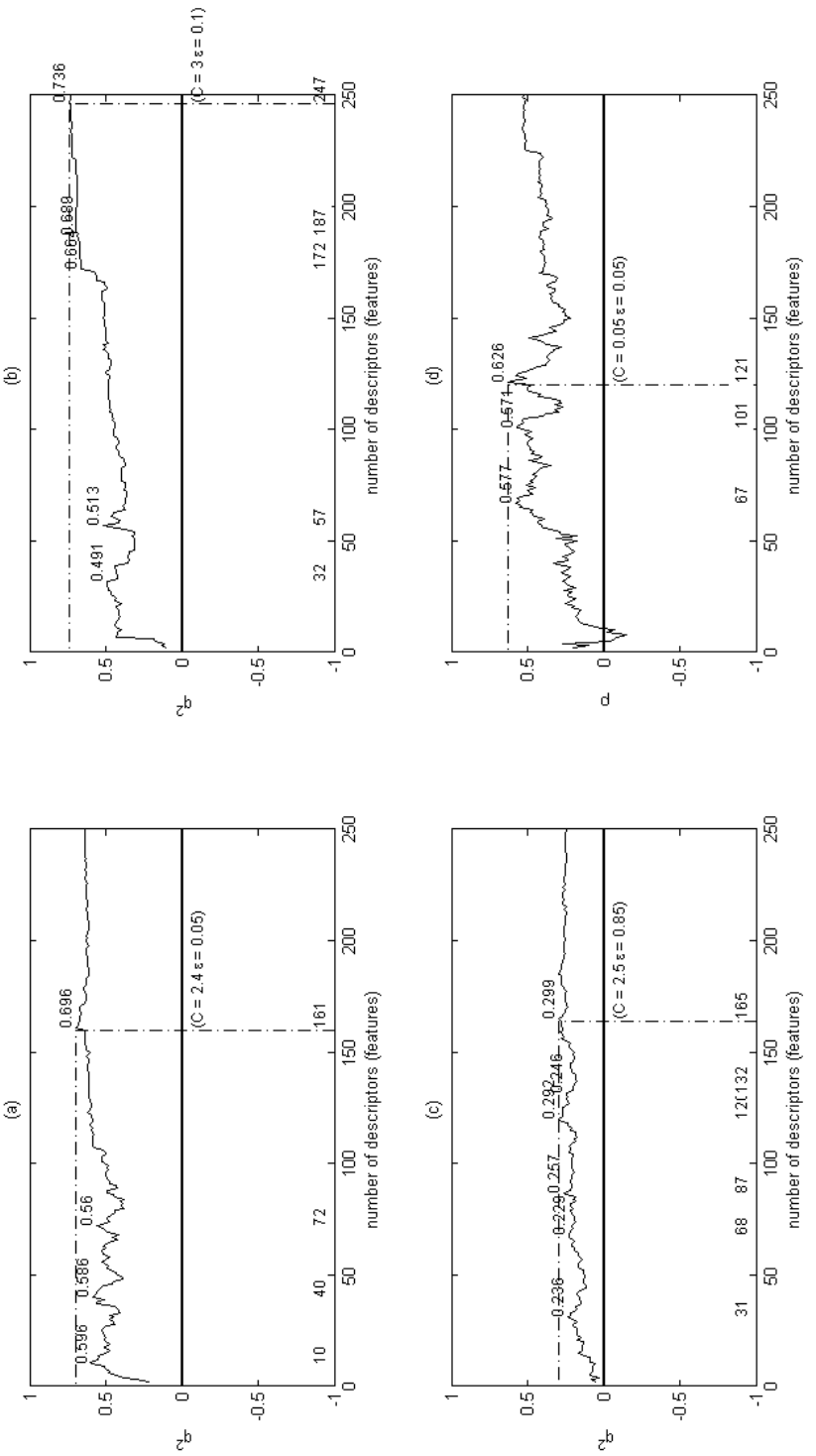


FIGURE 5.8: The performance of 7-rule fuzzy model based on the number of descriptors. a) Task 1: Graph shows distinct peaks when the number of descriptors are 10, 40, 72 and reaches highest peak at 161 with the SVR parameters (C = 2.4 and ϵ = 0.05). b) Task 2: Graph shows distinct peaks when the number of descriptors are 32, 57, 172, 188 and reaches highest peak at 247 with the SVR parameters (C = 3.0 and ϵ = 0.1). c) Task 3: Graph shows distinct peaks when the number of descriptors are 31, 68, 87, 120, 132 and reaches highest peak at 165 with the SVR parameters (C = 2.5 and ϵ = 0.85). d) Task 4: Graph shows distinct peaks when the number of descriptors are 67, 101 and reaches highest peak at 121 with the SVR parameters (C = 0.05 and ϵ = 0.05).

TABLE 5.4: Top most frequent amino acid features selected for the optimal model of Task 1 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	481	7	1	1	1	0	1	0	1	1	1
2	302	6	0	1	1	0	1	1	1	0	1
3	367	6	1	1	0	0	1	1	0	1	1
4	31	5	0	0	1	1	0	1	1	1	0
5	613	5	1	1	0	0	0	1	1	0	1
6	259	4	0	1	0	1	0	1	0	1	0
7	359	4	0	0	1	1	0	0	1	1	0
8	400	4	0	1	0	1	0	0	0	1	1

TABLE 5.5: Top most frequent amino acid features selected for the optimal model of Task 2 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	364	7	1	1	0	1	1	1	1	1
2	31	6	1	1	1	1	1	0	0	1
3	379	6	1	0	0	1	1	1	1	1
4	400	6	1	1	0	1	0	1	1	1
5	476	6	1	0	0	1	1	1	1	1
6	30	5	1	0	1	1	0	0	1	1
7	235	5	0	1	1	1	1	0	1	0
8	302	5	0	1	1	1	0	0	1	1
9	380	5	1	0	0	0	1	1	1	1
10	386	5	0	1	1	1	1	0	1	0
11	609	5	1	1	0	1	1	1	0	0

TABLE 5.6: Top most frequent amino acid features selected for the optimal model of Task 3 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	110	4	0	1	0	1	0	1	0	0	1
2	338	4	0	0	0	1	0	1	1	1	0
3	376	4	0	0	0	1	0	1	1	1	0
4	405	4	1	1	1	0	0	0	1	0	0
5	25	3	0	0	1	1	0	0	0	1	0
6	88	3	0	0	1	1	0	1	0	0	0
7	220	3	0	0	0	1	0	0	1	1	0
8	221	3	1	0	0	0	0	1	0	1	0
9	232	3	0	1	0	1	0	0	0	1	0
10	296	3	1	0	0	1	0	0	0	1	0
11	299	3	0	0	0	0	1	1	0	1	0
12	345	3	0	0	0	0	0	1	1	1	0
13	349	3	0	0	1	0	1	0	0	1	0
14	367	3	1	0	0	0	0	0	1	1	0
15	373	3	1	0	0	0	0	1	0	1	0
16	400	3	1	0	0	0	0	0	1	1	0
17	452	3	1	0	0	1	1	0	0	0	0
18	455	3	0	0	1	1	0	0	0	1	0
19	481	3	0	0	0	0	1	0	1	1	0

TABLE 5.7: Top most frequent amino acid features selected for the optimal model of Task 4 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	306	4	0	0	0	1	0	1	1	1	0
2	338	4	0	0	0	1	0	1	1	1	0
3	110	3	0	1	0	0	0	1	0	0	1
4	125	3	0	0	0	0	1	1	0	1	0
5	221	3	1	0	0	0	0	1	0	1	0
6	232	3	0	1	0	1	0	0	0	1	0
7	251	3	0	0	0	1	0	0	1	0	1
8	373	3	1	0	0	0	0	1	0	1	0
9	405	3	1	1	1	0	0	0	0	0	0
10	420	3	1	0	0	0	0	1	1	0	0

TABLE 5.8: Prediction results of the proposed model for each rule-base.

Number of Rules (Clusters)	Task 1 q^2 (features)	Task 2 q^2 (features)	Task 3 q^2 (features)	Task 4 ρ (features)
	0.692 (161)	0.671 (172)	0.236 (31)	0.598 (101)
2	0.692 (161)	0.739 (246)	0.299 (165)	**0.643 (141)**
	0.693 (161)	0.669 (176)	0.236 (31)	0.594 (101)
3	0.693 (161)	**0.743 (247)**	**0.310 (172)**	0.638 (141)
	0.693 (161)	0.671 (172)	0.236 (31)	0.587 (101)
4	0.693 (161)	**0.743 (247)**	0.299 (165)	**0.643 (141)**
	0.694 (161)	0.670 (172)	0.236 (31)	0.573 (67)
5	0.694 (161)	**0.743 (247)**	0.299 (165)	0.639 (141)
	0.695 (161)	0.668 (172)	0.236 (31)	0.582 (67)
6	0.695 (161)	0.740 (247)	0.299 (165)	0.628 (141)
	0.696 (161)	0.664 (172)	0.236 (31)	0.577 (67)
7	**0.696 (161)**	0.736 (247)	0.299 (165)	0.626 (121)

TABLE 5.9: SVR prediction results compared to the results of other SVR-based methods presented in the literature.

Group		Task 1	Task 2	Task 3	Task 4
Performance Measures		q^2	q^2	q^2	ρ
Gavin Cawley	[285]	0.677	0.305	-0.001	N/A
Liao Quan	[285]	0.601	N/A	N/A	N/A
Scott Oloff	[285]	0.586	0.363	N/A	N/A
Reiji Teramoto	[285]	0.374	0.401	0.154	0.565
Joao Aires-de-Sousa	[285]	-0.298	N/A	N/A	N/A
WTD-BBO-SVM	[292]	0.682	0.639	0.232	N/A
SVR		0.625	**0.741**	**0.299**	**0.575**
Improvement		-	15.9%	28.8%	1.7%

TABLE 5.10: Prediction results of the proposed model compared to the results found in literature.

Methods		Task 1	Task 2	Task 3	Task 4
Performance Measures		q^2	q^2	q^2	ρ
Number of Participants		14	10	7	6
First	[285]	0.677	0.735	0.236	0.593
Second	[285]	0.626	0.612	0.201	0.565
Third	[285]	0.615	0.455	0.154	0.472
L1 Regularization	[81]	0.667	0.642	0.205	0.548
L1, L2 Regularization	[81]	**0.691**	**0.668**	0.131	**0.586**
KPLS exponential	[291]	**0.691**	0.590	0.219	N/A
WTD-BBO-SVM	[292]	0.682	0.639	**0.232**	N/A
WT-BBO-RF	[292]	0.661	0.607	0.208	N/A
TSK-SVR I		**0.696**	**0.743**	**0.310**	**0.643**
Improvement		0.7%	11.2%	33.6%	9.7%

5.3.2 Cross-Validated Peptide Binding Affinity Prediction

The proposed model applied to two different prediction cases similar to cases studied in the literature for comparison purposes: entire data set prediction and leave-one-out cross validated correlation coefficient prediction. For each rule-base (rules that range between two and five), feature selection was carried out to reduce the number of features.

For the entire data set prediction as shown in Table 5.14, it can be seen that two different measures were used to observe their influence on the prediction error. The prediction results are comparatively better than those of the studies presented in [286], [25] and [306] for MHC alleles H2-Db and H2-Kb. The predictive performances have been improved by 7.9% (q^2) and 17.6% (AR) for the H2-Db allele; and 14.6% (q^2) and 10.9% (AR) for the H2-Kb allele. There is no improvement gain obtained for the H2-Kk allele. The optimal parameters for the MHC alleles using the q^2 measure are found to be: $C = 75.0$, $\epsilon = 0.20$ for H2-Db allele; $C = 25.0$, $\epsilon = 0.50$ for H2-Kb allele; $C = 18.5$, $\epsilon = 0.20$ for H2-Kk allele. The models contained 30, 25 and 62 features for each MHC allele, respectively. The average residual (AR) measure values of the proposed model are: $C = 9.75$, $\epsilon = 0.05$ for allele H2-Db; $C = 9.65$, $\epsilon = 0.05$ for allele H2-Kb; and $C = 7.5$, $\epsilon = 0.05$ for allele H2-Kk. The final and refined models contained 39, 24 and 22 features, respectively. In order to further explain the results for the entire data set prediction, the construction of correlation diagram (Fig. 5.9) for each allele data set is used to illustrate the relationship between the experimentally measured and predicted pIC50 values. When the performance is perfect, the correlation diagram shows a straight line along the 45° diagonal. A good quality of prediction performance can be obtained when the data samples are mainly distributed along the 45° diagonal. The divergence in the line is caused by the prediction error between the measured and the predicted pIC50 values.

In addition, each model was evaluated by using leave-one-out cross validation (LOO-CV) using the cross-validated correlation coefficient. This will allow an independent predictive assessment as compared to the assessment carried out using the entire data set. As the compared methods presented in the literature did not report average residual measure for the LOO-CV experiments, this assessment was excluded from the calculations. The additive method recognized 6 outliers for H2-Db, 7 outliers for H2-Kb and 2 outliers for H2-Kk. Nevertheless, SVRMHC method did not recognize any outliers for

the H2-Kk and obtained a result much better than the additive method. These outliers are removed prior to LOO-CV calculations for the additive and SVRMHC methods for the H2-Db and H2-Kb. For the H2-Kk, however, additive method excludes two outliers whereas SVRMHC method does not exclude any outliers. The same outliers are also excluded (except for H2-Kk similar to SVRMHC method) from the proposed models during the LOO-CV calculations in order to perform a consistent comparison. The optimal parameters for the MHC alleles using the q^2 measure are found to be: $C = 0.45$, $\epsilon = 0.05$ for H2-Db allele; $C = 0.25$, $\epsilon = 0.15$ for H2-Kb allele; $C = 3.10$, $\epsilon = 0.05$ for H2-Kk allele. The models contained 34, 32 and 21 features for each MHC allele, respectively. It should be noted that selected features are highly dependent on their data sets. Approximately 0.5% of the features are adequate for finding the optimal models. As shown in Table 5.15 the proposed models yielded LOO-CV q^2 values of 0.462, 0.490, and 0.729 which are higher predictive accuracy than the additive and SVRMHC methods for each MHC allele, respectively. The predictive performance for Tasks H2-Db, H2-Kb, and H2-Kk have been improved by 1.32%, 0.82%, and 1.11% to the best model presented in the literature, respectively. The overall improvement gain for all tasks is found to be 1.08%.

The amino acid features that contributed most to the efficiency of the proposed models are given in Table 5.11 - Table 5.13. Only the proposed models found using leave-one-out cross validation (LOO-CV) take into consideration as they allow an independent predictive assessment as compared to the assessment carried out using the entire data set. For H2-Db, five amino acid features contributed to the output in two separate locations. The amino acid feature numbered with 18 (Spin-spin coupling constants), 27 (The number of atoms in the side chain), 88 (Positive charge), 481 (Hydrophobicity coefficient in reversed phase high performance liquid chromatography), 520 (Unknown) contributed highest as they are represented in two separate locations on each of the nona-peptide within the data set. For H2-Kb, one amino acid feature contributed to the output in two separate locations. The amino acid feature numbered with 71 (Direction of hydrophobic moment) contributed highest as it is represented in two separate locations on each of the octa-peptide within the data set. For H2-Kk, three amino acid features contributed to the output in two separate locations. The amino acid feature numbered with 29 (The number of bonds in the longest chain), 88 (Positive charge), 565 (Unknown)

contributed highest as they are represented in two separate locations on each of the octapeptide within the data set. The amino acid feature numbered with 88 (Positive charge) appeared in H2-Db and H2-Kk as a common feature with location occurrences of 2 and 2, respectively. Therefore, the positive charge of an amino acid is considered as one of the highly discriminating feature in these data sets. This finding suggests that positive charge is important in mediating the binding process between the peptide and MHC molecule in this data set. It is reported that positively charged amino acids can play important role for the transmembrane domains of reduced folate carrier [307]. The full descriptions of amino acid features can be found in Appendix A.

It should be noted that our literature search appears to indicate that these data sets have been understudied due to their complexity, therefore not many papers other than the cited ones seem to have appeared in the literature [25], [286], [306]. The cross-validated results suggest that a better descriptive power has been achieved over the unseen data indicating better generalisation ability of the proposed hybrid method. In addition, the incorporation of fuzzy system with SVR has enabled to improve SVR and consequently resulting in a better modelling of uncertainty even the model can only use small sample size being the nature of peptide data. As stated above, the fuzzy if-then rule set proposed suggests promising results.

TABLE 5.11: Top most frequent amino acid features selected for the optimal model of H2-Db and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	18	2	0	0	0	1	0	0	1	0	0
2	27	2	0	0	1	0	0	0	0	0	1
3	88	2	0	1	0	1	0	0	0	0	0
4	481	2	0	0	0	0	0	0	1	0	1
5	520	2	0	0	0	1	0	0	1	0	0

TABLE 5.12: Top most frequent amino acid features selected for the optimal model of H2-Kb and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	71	2	0	1	0	0	1	0	0	0

TABLE 5.13: Top most frequent amino acid features selected for the optimal model of H2-Kk and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	29	2	0	1	0	0	0	0	0	1
2	88	2	1	0	0	0	0	0	0	1
3	565	2	0	0	0	0	1	0	1	0

TABLE 5.14: Entire data set prediction results of the mouse class I MHC alleles.

	Allele			Allele		
	$H2-D^b$	$H2-K^b$	$H2-K^k$	$H2-D^b$	$H2-K^b$	$H2-K^k$
Methods	q^2	q^2	q^2	AR	AR	AR
Additive [286]	0.602	0.370	0.849	0.403	0.443	0.178
SVRMHC [25]	0.749	0.568	0.973	0.170	0.382	0.039
RVMMHC-1 [306]	0.840	0.664	0.980	0.297	0.527	0.125
RVMMHC-2 [306]	0.845	0.691	0.962	0.316	0.489	0.173
TSK-SVR I	**0.912**	**0.792**	0.912	**0.140**	**0.340**	0.193
Improvement	7.93%	14.62%	-	17.65%	10.99%	-

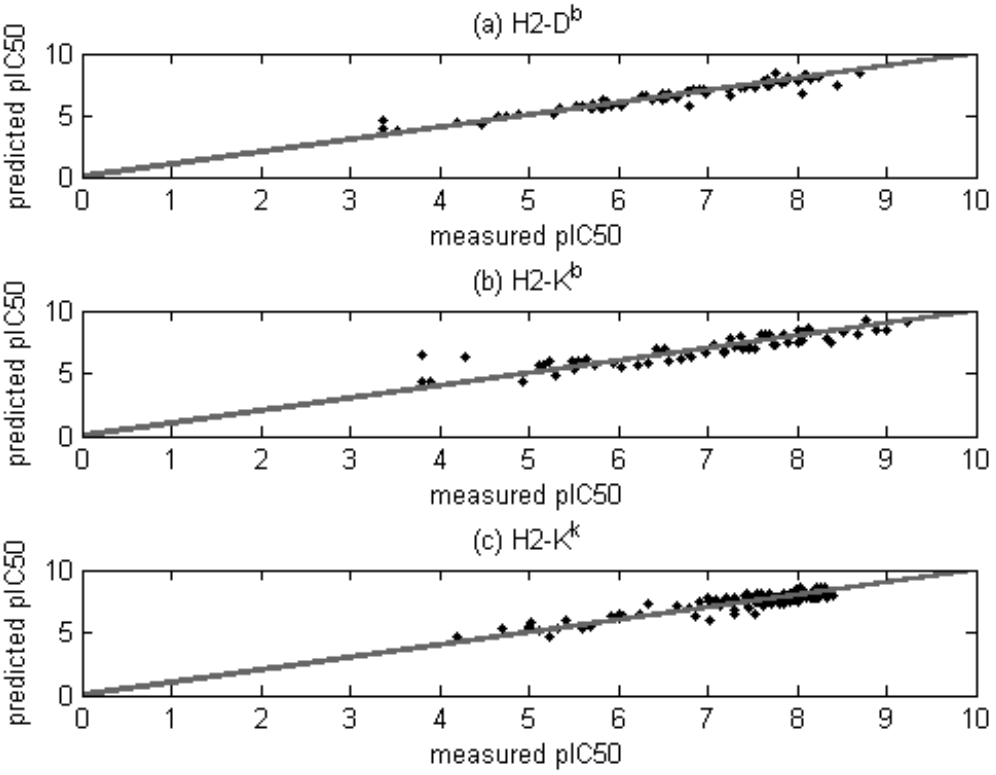


FIGURE 5.9: Correlation diagrams of the prediction performance for mouse class I MHC alleles. a) H2-Db b) H2-Kb c) H2-Kk

TABLE 5.15: Leave-one-out cross validated correlation coefficient (q^2) prediction results of the mouse class I MHC alleles.

		Allele		
		$H2-D^b$	$H2-K^b$	$H2-K^k$
Methods		q^2	q^2	q^2
Additive	[286]	0.401	0.454	0.456
SVRMHC	[25]	0.456	0.486	0.721
TSK-SVR I		**0.462**	**0.490**	**0.729**
Improvement		1.32%	0.82%	1.11%

5.4 Conclusions

In this chapter, a hybrid system that has helped to improve the predictive ability of fuzzy system significantly with the aid of support-based vector method was developed. The proposed method demonstrated with the successful applications in the prediction of peptide target value being regarded as one of the difficult modelling problems in bioinformatics. Two major points were identified. First, SVR is enhanced by adding the fuzziness concept. Second, TSK-FS is benefited from SVR-based training. The SVR-based experiments were carried out for four different peptide affinity data sets and three mouse class I MHC alleles. The experimental results evidently highlight the strength of the proposed hybrid method which yielded comparatively better results among the recently published results. Predictive performances have been improved as much as 33.6% for the first group of data sets and 1.32% for the second group of data sets. Apart from improving the prediction accuracy, this research study has also identified amino acid features "Polarity", "Positive charge", "Hydrophobicity coefficient", and "Zimm-Bragg parameter" being the highly discriminating features in the peptide binding affinity data sets.

Chapter 6

Quantitative Prediction of Peptide Binding Affinity with SVR-based Interval Type-2 Fuzzy System

6.1 Introduction

Peptide binding plays important roles in the immune system and helps us to understand the mechanisms of protein-peptide interactions. One of the most important aspects of the binding of peptides is the prediction of protein-peptide binding affinity with applications to design of drugs. Empirical evaluation of the binding affinity is unfeasible as there are huge number of potential binding peptides even for a particular major histocompatibility complex molecule. Furthermore, it requires laboratory experiments that are costly and time consuming. The use of computational methods are inevitable to support empirical methods in order to determine the binding and its affinity in a quicker manner. Predictive models help approximate computation of the tendency and strength of the bindings and serve as essential time saving tools.

Fuzzy systems can be used in modelling of uncertain systems and imprecise knowledge very similar to human reasoning [168]. Expert knowledge traditionally is the main source when designing a fuzzy system. Nevertheless it is difficult to find the human experts

when they are needed. Moreover, it is infeasible to ask them repeatedly when modifications are required. On the contrary, the necessities of real-life applications often require to adapt the modifications that may occur in the environment. One of the generally used fuzzy systems is the Takagi-Sugeno-Kang fuzzy system [6], [7]. It can model complex systems and can be enhanced with the cooperation of learning methods. In this regard, consequent parameters of a TSK fuzzy system can be obtained using the least square estimation. As the training error is minimised during the least squares estimation the model can lead to overfitting. Support Vector Regression is an acceptable alternative regression estimation method to the least squares and can ensure generalisation of underlying model.

In order to improve the accuracy of the fuzzy models and minimize the affects of uncertainties, type-2 fuzzy systems are used [179]. Type-2 fuzzy sets assist in knowledge representation by the use of linguistic grades of membership and improve the inference of type-1 fuzzy sets [308]. The computations of type-2 fuzzy sets are complex. In order to ease these computations interval type-2 sets can be used [190]. IT2 fuzzy sets are often much more practical to manage than the general type-2 fuzzy sets. One of the advantages of using IT2 fuzzy sets is that the computations can be implemented using type-1 fuzzy sets [187]. Similar to the defuzzification process in type-1 fuzzy systems, type-2 fuzzy systems use type-reduction process in order to find a type-1 set [309]. Karnik-Mendel algorithms are the widely used type-reduction algorithms and compute the centroid of IT2 fuzzy sets in order to find a type-reduced set [310]. The iterative nature of the KM algorithms often leads a computational cost which in turn results inefficiency when they are used in fuzzy logic control systems [311].

A hybrid learning system that incorporates the Type-2 TSK fuzzy system with SVR and clustering methods are proposed in this chapter in order to built a robust fuzzy predictive model. The consequent parameters are obtained by SVR whereas antecedent parameters of the fuzzy system are obtained using clustering methods. Recently, a general framework that integrates type-2 fuzzy system with the SVR-based method has been presented [312]. In order to address the computational cost of a type-reduction process, our approach used a different inference engine in which type-reduction is not necessary. To initialize the parameters of IT2 fuzzy sets, a novel clustering concept is developed. This clustering approach is based on the overlapping concept.

In the next section, the proposed type-2 fuzzy system is described in detail. In Section 6.3 the results of the binding affinity problem are presented and discussed. Finally, Section 6.4 draws the conclusions of this chapter.

6.2 Materials and Methods

In this section, the proposed type-2 fuzzy system is described in the following subsections: IT2-TSK A2-C0 Fuzzy System (6.2.1), Type Reduction and Defuzzification (6.2.2), Generating Fuzzy System with Overlapping Clustering Concept (6.2.3), SVR-based IT2-TSK Fuzzy System (6.2.4).

6.2.1 IT2-TSK A2-C0 Fuzzy System

Takagi-Sugeno-Kang model is one of the widely used fuzzy systems. This model structure presents the design of consequent parameters using a least squares method. Moreover, model structure is extended in such a way that it can identify both premise and consequent part of the fuzzy system.

The rule-base of the IT2-TSK A2-C0 model with r rules can be expressed as:

$$\begin{aligned} &R_i : \text{IF } x_1 \text{ is } \tilde{A}_{1i} \text{ AND } x_2 \text{ is } \tilde{A}_{2i} \text{ ... AND } x_n \text{ is } \tilde{A}_{ni} \\ &\text{THEN } y_i \;=\; a_{0i} + a_{1i}x_1 + ... + a_{ni}x_n \end{aligned} \tag{6.1}$$

where n are the input variables (x_1, x_2, ..., x_n); and $\tilde{A}_{ni}$ is an interval type-2 fuzzy set for the variable n and rule r, generally represented by a membership function; and y_i is a linear function in the consequent part; and a_0, a_1, a_2, ..., a_n are the coefficients of input parameters. As the model structure is A2-C0, the coefficients of the consequent are crisp numbers.

The antecedent part involves IT2 fuzzy sets where the uncertainty is modeled. The firing strengths of IT2-TSK are determined by using the t-norm operator and can be

calculated as:

$$\underline{f_i} = \prod_{j=1}^{n} \underline{\mu}(x_j) \tag{6.2}$$

$$\overline{f_i} = \prod_{j=1}^{n} \overline{\mu}(x_j) \tag{6.3}$$

where $\underline{f_i}$ and $\overline{f_i}$ are the lower and upper firing strengths; $\underline{\mu}(x_j)$ and $\overline{\mu}(x_j)$ are the upper and lower degree of memberships for input variable x_j; respectively, and $\prod$ denotes the product t-norm operation.

6.2.2 Type Reduction and Defuzzification

Interval type-2 fuzzy systems are often used to model and minimise the effects of uncertainties in fuzzy systems [186]. Type-reduction process is an important step in IT2-FS. This process enables to reduce a type-2 fuzzy set into a type-1 fuzzy set. Karnik-Mendel algorithm is a widely used type-reduction method that can compute the left and right end points required for IT2 fuzzy set [191]. Then these end points are used to calculate the final output. Due to the high-computational cost of iterative KM algorithms, alternative type-reduction algorithms that are faster in computation and have closed form expressions have been proposed recently in the literature. Some of the computationally effective alternative type-reduction algorithms, many of them are for the defuzzification of Mamdani IT2 fuzzy logic systems, are Liang-Mendel Unnormalised Method [313], Wu-Mendel Uncertainty Bounds Method [314], Coupland-John Geometric Method [315], Greenfield-Chiclana-Coupland-John Collapsing Method [316], Nie-Tan Method [317].

Wu-Mendel's uncertain boundary method (WM) is an alternative for finding the overall output Y. This type-reduction method benefits from uncertainty bounds for IT2-FS in order to decrease the computational load. WM method uses four centroids ($\underline{y}_l$, $\underline{y}_r$, $\overline{y}_l$, $\overline{y}_r$) which are the left and right end points of the centroid of the consequent IT2-FS.

One important note about the WM method is that the overall output can be calculated without having to perform type-reduction.

$$Y_{\mathrm{WM}} = 1/2\ [\ \frac{\underline{y}_l + \overline{y}_l}{2} + \frac{\underline{y}_r + \overline{y}_r}{2}\ ] \tag{6.4}$$

One main drawback of WM method is that there is no systematically designed for IT2 fuzzy control systems and stability analysis of the output equations reported unsuccessful. Biglarbegian-Melek-Mendel (BMM) proposed a new inference engine that designs the parameters of IT2-TSK [318]. This method has a closed mathematical form and conditions required for the stability of IT2-TSK. However, BMM method gets more of its theoretical background based on WM's method and suggested their new inference method as described in the following. BMM introduced a new inference engine as:

$$Y_{\mathrm{BMM}} = q\frac{\sum_{i=1}^{r} \overline{f_i}\ y_i}{\sum_{i=1}^{r} \overline{f_i}} + p\frac{\sum_{i=1}^{r} \underline{f_i}\ y_i}{\sum_{i=1}^{r} \underline{f_i}} \tag{6.5}$$

where q and p are the design parameters to weight the lower ($\underline{f_i}$) and upper ($\overline{f_i}$) firing strengths for each rule, respectively (if $r = 1$, then $q + p = 1$). These parameters are required to be optimised for the robustness of the fuzzy system. The rule outputs denoted by y_i are not required to be sorted in BMM type reduction.

6.2.3 Generating Fuzzy System with Overlapping Clustering Concept

In the proposed work of Sugeno and Yasukawa [319], fuzzy clustering is used to identify the structure and parameters of a fuzzy model. This work also classifies the identification process into two kinds and describes each of them thoroughly. These are structure identification and parameter identification in fuzzy modelling. Finding the input variables from the possible input space and determining the number of rules are the main concerns in structure identification. Parameter identification, however is mostly concerned with finding parameter values of the fuzzy model. These parameter values, in the case of premise parameter identification, can be of a non-linear nature, are used to form

the membership functions which characterise fuzzy sets. Later, to ease structure identification process, sample probability distributions were suggested in order to identify parameters of membership functions of input variables using the centres of cluster-like regions [202].

Although structure identification and parameter identification help to better deal with a complex system like type-1 fuzzy systems. In the case of type-2 fuzzy systems, the matter becomes even more complex. To our best knowledge, there is no accepted method in the literature for the parameter initialization of a type-2 fuzzy system. In the case of type-1 fuzzy system, the knowledge obtained from a domain expert can be used in this fashion. However, in the absence of a domain expert, a common practice is to use uniform fuzzy partitioning based on a number of labels for each feature [320], [321]. It is obvious that in the case of a high-dimensional feature space this approach will not do. Because the feature size is large, the rule-base is formed of huge number of rules. Consequently, this leads to the curse of dimensionality problem, which one would like to particularly avoid. So the grid-partitioning is omitted from the efforts of type-2 fuzzy system premise parameter initialization. In the literature, one effort found in this fashion so far is to derive the lower MF from the given upper MF [205]. Above all, it is considered that for IT2 fuzzy systems, arbitrary initialization of MF parameters is the common practice. After the arbitrary initialization, a learning method is used for finding the optimum parameters. As a result, the model structure of T2-TSK is often a difficult task. A novel method is therefore developed based on the overlapping concept in order to ease this tedious task. IT2 premise parameters consist lower and upper membership functions. In this method upper membership functions are identified using the clustering approach similar to strategy discussed to identify the membership functions of the type-1 fuzzy system. The lower membership functions on the other hand are identified from the overlapping regions among the clusters.

The clustering method introduced in this section aims for overcoming the difficulties of parameter identification process in a type-2 fuzzy system. This method assumes the overlapping regions between the clusters may contain uncertain parts that could be useful to take into consideration in the design process of an interval type-2 fuzzy set. Upper membership function parameters of an IT2-FS is obtained using the chosen clustering method. This chosen clustering method can be any clustering method as long as the clusters provide the statistical information defining their characteristic. This work used

common clustering methods like k-means, fuzzy c-Means and hierarchical clustering in the experiments and made a performance comparison among them.

The overlapping clustering process is similar to the approach taken in the previous chapter when determining the membership function parameters of a type-1 fuzzy set. But it differs in that an IT2-FS requires its lower membership function to be defined. The projection of these overlapping regions defines new end points for each cluster. A cluster which is located on the left wing of the designated cluster may define its left end point and the other cluster on its right wing may define its right end point. The use of these new points obtained through the projecting of clusters into 1D representations and with the addition of cluster center, the parameters of lower membership function would be obtained. The overlapping clustering concept is illustrated in Fig. 6.1.

The overlapping clustering concept is comprised of the following main steps:

Step 1: Decide the number of clusters for the partitioning clustering methods or cut-off point for the hierarchical clustering methods.

Step 2: Do the cluster analysis for the chosen clustering method (e.g. HCM, FCM, Hierarchical Clustering). Data samples are assigned to each cluster at the end of the clustering process. Get the statistical values of each generated partition. These statistical values (e.g. min, max, mean, standard deviation, variance) of each partition can be used to determine values of the parameters of a membership function in fuzzy modelling. Note that in the case of the standard deviation equals to zero for the generated partitions, set the standard deviation to a small but non-zero value. The purpose is to avoid clusters having zero standard deviation and ensuring the membership functions (e.g. Gaussian membership function) work properly during the fuzzification stage.

Step 3: Overlapping clustering concept can be applied to any type of membership functions however in this explanation for the sake of simplicity and clarification of example it is explained for two different kind/type of membership functions mainly triangular membership functions and Gaussian membership functions. Gaussian membership function depends on two parameters. They are standard deviation and mean. Triangular membership function depends on three parameters. These are the left, right points and mean. After the cluster analysis, get these aforementioned statistical values of all partitions.

Step 4: Continue the same process in steps 5-11 for each single-feature in the feature-set. Get the pair (statistical value, designated partition) for each single-feature.

Step 5: Set the upper left point at the value obtained from the pair (min, partition). Set the upper center by the value obtained from the pair (mean, partition). Set the upper right point by the value obtained from the pair (max, partition). These values characterize the upper membership function.

Step 6: Initialize the parameters of the lower membership function by using values that characterize the upper membership function. Set the lower left point to upper right point. Set the lower centre to upper centre. Set the lower right point to upper right point.

Step 7: Setting the lower left point: Get the min, mean, max values of all the partitions. Find the lower left point by searching all these statistical values of the partitions and obtain the value in that the value shall be in the interval [leftpoint, mean] of upper statistical values in the other partitions. If there is more than one value found in the search process. Get the closest value to the upper left point.

Step 8: Setting the upper left point: Get the min, mean, max values of all the partitions. Find the upper left point by searching the all the statistical values and obtain the value in that the value shall be in the interval [rightpoint, mean] of upper statistical values in the other partitions. If there is more than one value found in the search process. Get the closest value to the upper right point.

Step 9: Setting the lower centre: As this is a membership function with fixed mean and uncertain standard deviations, the upper and lower centres remain the same.

Step 10: Generate upper triangular/Gaussian membership function using the upper left, right end points and centre.

Step 11: As absolute lengths of lower left point and lower right point from the centre are not the same. Generate two lower triangular/Gaussian membership functions representing each end points. Form a non-uniform lower triangular/Gaussian membership function taking left wing from one triangular/Gaussian membership function and take the right wing from the other. Ensure the lower membership function generated has a non-zero standard deviation. In the case of it equals to zero, set the standard deviation of lower membership function to a small but non-zero value.

Step 12: Either use triangular/Gaussian membership functions or convert them to other membership functions (e.g. trapezoidal membership function) to use in fuzzy modelling.

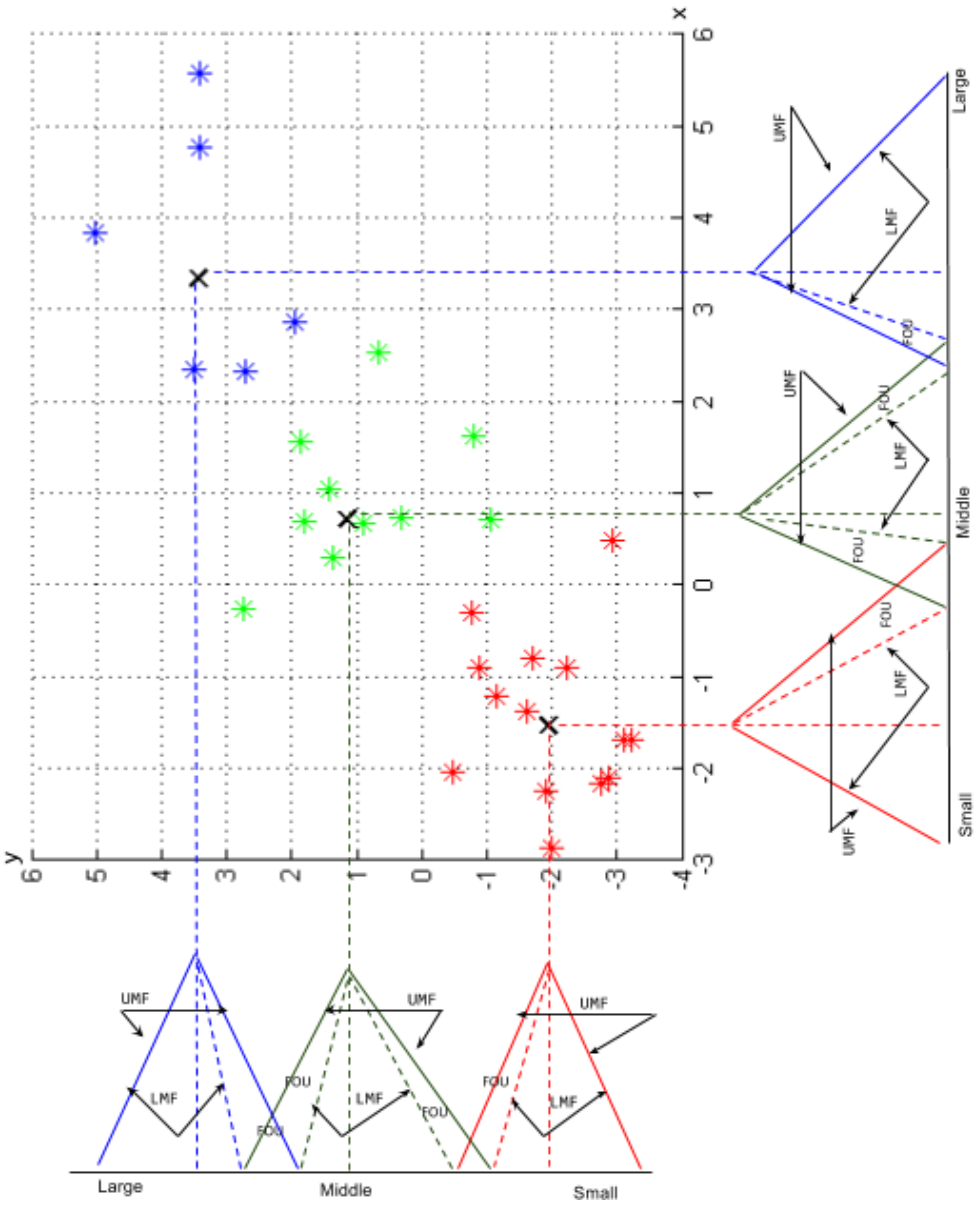


FIGURE 6.1: Illustration of determination of the initial values of the parameters of triangular membership functions of IT2 lower and upper membership functions derived from overlapping clustering concept. UMF: upper membership function; LMF: lower membership function. The bounded region is called a footprint of uncertainty.

6.2.4 SVR-based IT2-TSK Fuzzy System

SVM is a powerful method based on the statistical learning theory, or VC theory and this theory uses the characteristics of learning machines that can lead a good generalisation for the unseen data [157]. In the case of regression estimations SVM can be referred as SVR. Given the training data as in the form of data pairs (x,y), SVM learning algorithm finds the function h that tolerates errors up to ϵ from the expected values of the targets y while ensuring the function as flat as possible. This means that, the errors less than ϵ can be tolerated as long as the deviations are not greater than the ϵ value.

$$h(x) = w^T x + b. \tag{6.6}$$

where w and b denote the coefficients of the linear function. The flatness of the function can be ensured on the search of small w with ϵ precision. Nevertheless, to cope with the infeasible constraints of the optimisation problem, slack variables ξ^+, ξ^- can be used.

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\|w\|^2 + C\sum(\xi_+ + \xi_-) \\ &\text{subject to} \begin{cases} y - (w^T x + b) \leq \epsilon + \xi_+ \\ (w^T x + b) - y \leq \epsilon + \xi_- \\ (\xi_+, \xi_-) \geq 0 \end{cases} \end{aligned} \tag{6.7}$$

The parameter C is the trade-off between deviations from the ϵ could be tolerated and the flatness of linear function h that up to which value of w could be minimised most. This is ensured by the ϵ-insensitive loss function where deviations of the data samples outside the tolerated value of ϵ are penalised and contribute to the cost function. As corresponded to the SVM, the training instances that have the non-vanishing coefficients are chosen as the support vectors. Accordingly then, the weighted sum of the support vectors characterises the separating hyperplane which acceptably models the training data set.

Least-squares estimation is a simple and standard method used to find the values of the consequent parameters of TSK [6]. A potential substitution of this common method

can be the SVR concept with a linear kernel. Training data set along with their target outputs are given to SVR benefiting from the BMM inference engine accordingly after inputs are transformed as:

$$(q\overline{f_i} + p\underline{f_i},\ q\overline{f_i}x_{i1} + p\underline{f_i}x_{i1},\ ...,\ q\overline{f_i}x_{in} + p\underline{f_i}x_{in}) \tag{6.8}$$

where q and p are the design parameters that denote weight of the lower and upper firing strengths for each rule. These weight parameters are optimised using a grid search to provide the robustness of the fuzzy system. Accordingly then, the coefficients w and b which represent the weight vector of the SVR linear function are computed. Thus a support vector based Type-2 Takagi-Sugeno-Kang fuzzy system (TSK-SVR II) can be formulated as:

$$y_i'' = w_{0r} + \sum_{i=1}^{n}(w_{ir}x_i) \tag{6.9}$$

$$y'' = q\frac{\sum_{i=1}^{r}\overline{f_i}\ y_i}{\sum_{i=1}^{r}\overline{f_i}} + p\frac{\sum_{i=1}^{r}\underline{f_i}\ y_i}{\sum_{i=1}^{r}\underline{f_i}} + b \tag{6.10}$$

where y'' denotes the new output formulation representing TSK-SVR II. SV-based regression that is used to compute the values of the consequent parameters of the hybrid method, is implemented using LIBSVM software.

6.2.5 Predictive Modelling of Peptide Binding Affinity

This section presents the construction of SVR based interval type-2 TSK fuzzy models and identification of their parameters in the following steps. The SVR based interval type-2 fuzzy model (TSK-SVR II) shown in Fig. 6.2 is used for the prediction of peptide binding affinity data sets. The figure illustrates the stages of this fuzzy model aiming at predicting degree of peptide binding.

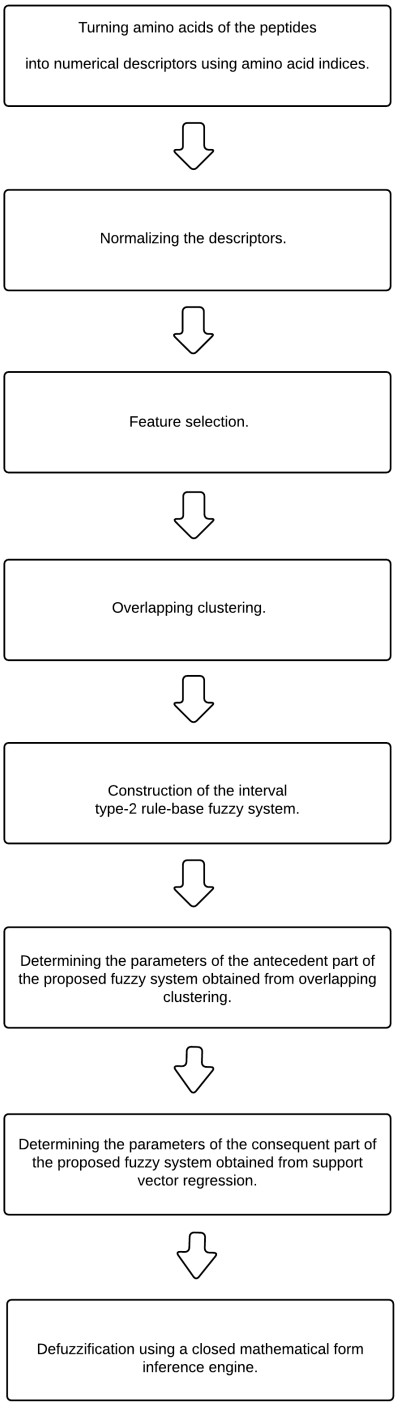


FIGURE 6.2: Stages of the SVR based interval type-2 TSK fuzzy model for the prediction of peptide binding affinity.

6.2.5.1 Preprocessing

The process starts with encoding a feature space from the available peptides using the amino acid descriptors. Each descriptor defines physico-chemical attribute values of 20 amino acids. For each amino acid location in the peptide, corresponding descriptor value for that amino acid is captured from the AA-scales. As described previously (Section 4.3), 643 descriptors have been used to define an amino acid for each location within the peptide. The constructed feature space varies according to the size of peptides. As peptides used in this research vary in size, mostly it would be 8 or 9, the number of features encoded becomes over five thousand features. This means that a high-dimensional feature set is used in order to carry out the process. After the completion of setting up the feature space, next step is the normalisation stage. Each feature converted to a real number between zero and one.

6.2.5.2 Feature Selection

This stage enables to ignore similar features leading to better descriptive features to be recognized. As a feature selection method, namely MCFS method [273], is used to filter the peptide data sets resulting in a reduced number of physico-chemical attributes. The subset of features found at the end of the feature reduction process is crucial to deal with what so called curse of dimensionality effect in prediction models. This effect drives the prediction models to become inefficient by demanding longer processing times and bigger memory sizes. However, MCFS method itself might suffer from the curse of dimensionality in the case of selection of high number of features. Hence, the number of features should be kept as low as possible.

6.2.5.3 Identifying Antecedent Parameters

Different than the type-1 fuzzy systems that require only one fuzzy set defined for each variable in the rule, interval type-2 fuzzy systems however require two fuzzy sets to describe an interval type-2 fuzzy set for each variable in the rule. An interval type-2 fuzzy set consists of lower and upper membership functions which are the boundaries of this type of fuzzy set. Each fuzzy set resides within these boundaries assumes a full membership value. To initialize the parameters of IT2 fuzzy sets, a novel clustering

concept is developed. This clustering approach is based on the overlapping concept. It takes into consideration that each single-variable is individually processed from the partitions generated in the hyperspace. This single input - single output scheme has partitions that overlaps each other. These regions are used as FOU. Overlapping concept can be used for any clustering method as long as the indices of which partition the data sample belongs to is provided. During the fuzzification stage, IT2 fuzzy sets for each variable in each rule are formed through the use of this novel strategy.

6.2.5.4 Identifying Consequent Parameters

For the fuzzy inference, a t-norm operation is used to find the firing strengths of each rule (both lower and upper firing strengths). In the type-reduction and defuzzification stage, a closed type-reduction strategy, namely BMM method, is followed. The firing strengths are combined with the design parameters of this method to weight the output of each rule, as described broadly in the relevant section.

The parameters of the consequent part for the TSK fuzzy systems are commonly initialized through the use of least squares. As our fuzzy model concerns IT2 fuzzy sets in its antecedent part, the consequent part is still type zero. So the similar approach as we used for finding parameters of a type-1 TSK fuzzy systems can be adopted here. Different from the least squares, our model uses support vector based regression in order to reveal the consequent parameters, contributing to better generalisation in the prediction process.

6.2.5.5 Searching for Optimal Parameters

As previously mentioned, the parameter to indicate number of clusters should be preset before the cluster analysis is performed. Therefore, silhouette graphs are obtained for two to seven clusters (Fig. 6.3 - Fig. 6.6) and for two to five clusters (Fig. 6.7 - Fig. 6.9) in order to reveal which groupings better represent the underlying data sets. Silhouette graphs suggest IT2-TSK fuzzy system can be constructed using only two rules with the reduced features. These rules are suffice for the proposed model to build a robust and interpretable fuzzy system for the high-dimensional data set by using relatively small number of data samples.

The structure of the IT2-TSK fuzzy system is constituted by automating the parameters of the antecedent and consequent parts. The optimal set of parameter values found at the end of grid-search are used. The values of the parameters of Gaussian membership functions that characterise each fuzzy set of the premise part were obtained by using clustering analysis such as k-means, fuzzy C-means, hierarchical clustering methods. The coefficients of linear functions of each rule for the consequent part were then identified using SVR. However, two more additional paremeters (q and p) needed to be optimised for the BMM method in the defuzzification stage of the model. These parameters are optimised using the grid-search while the SVR parameters remained constant as they are found at the end of intensive seeking process. The optimal parameters of linear kernel SVR (C and ϵ) and number of selected features along with the optimal design parameters of BMM method that weights the lower and upper firing strengths that resulted in best performance are given in Table 6.1 and Table 6.2.

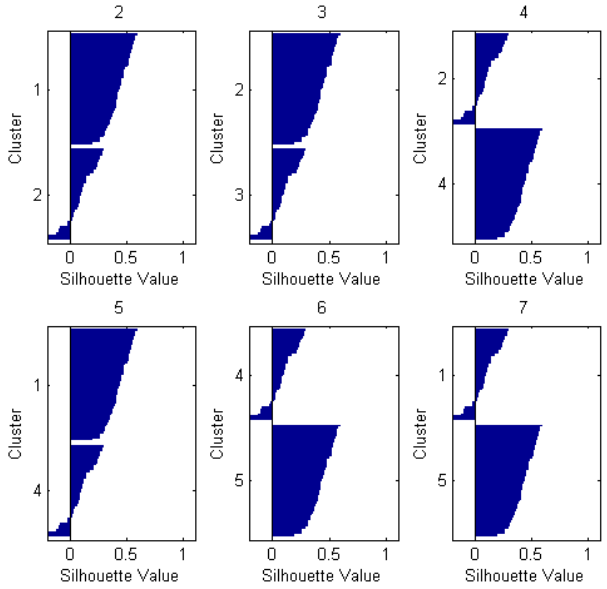


FIGURE 6.3: Silhouette values for different clusters for Task 1.

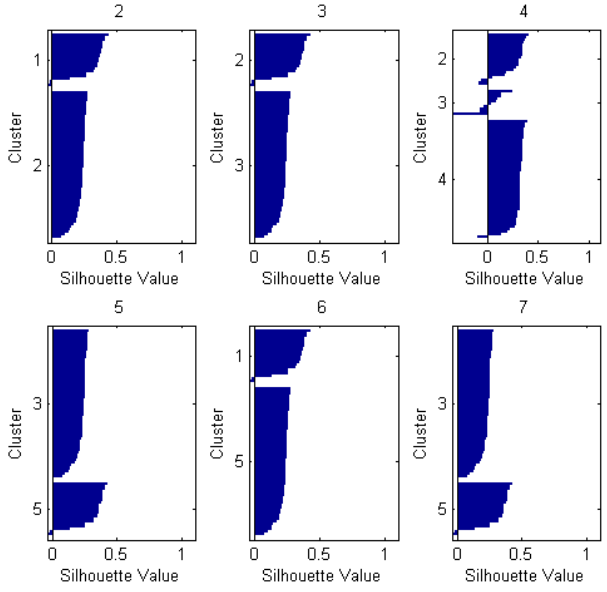


FIGURE 6.4: Silhouette values for different clusters for Task 2.

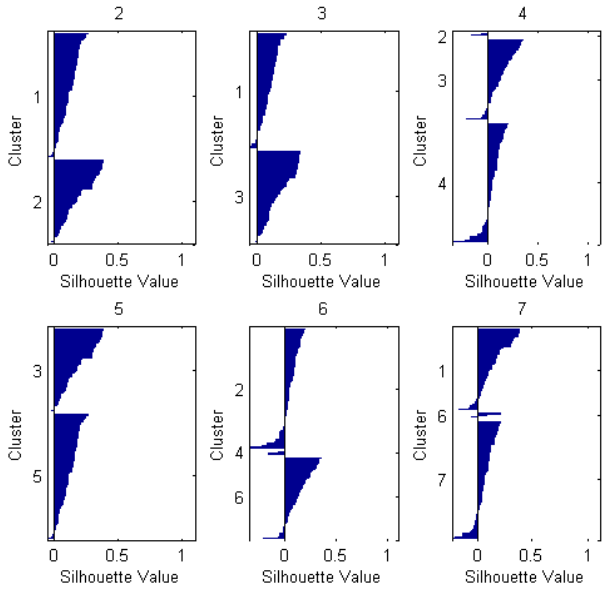


FIGURE 6.5: Silhouette values for different clusters for Task 3.

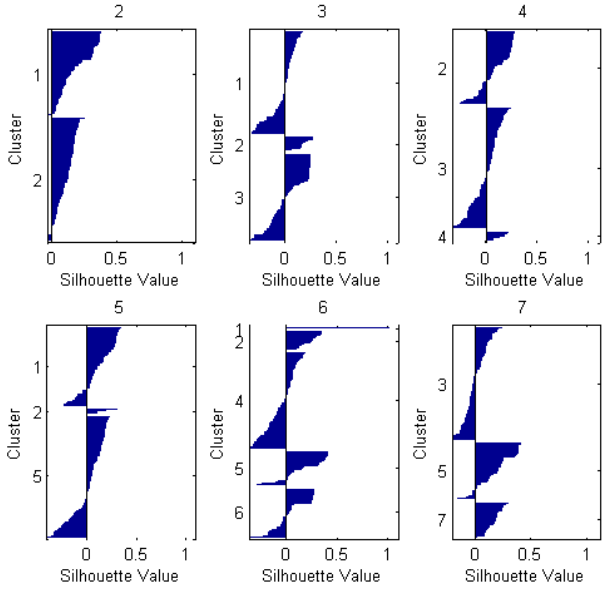


FIGURE 6.6: Silhouette values for different clusters for Task 4.

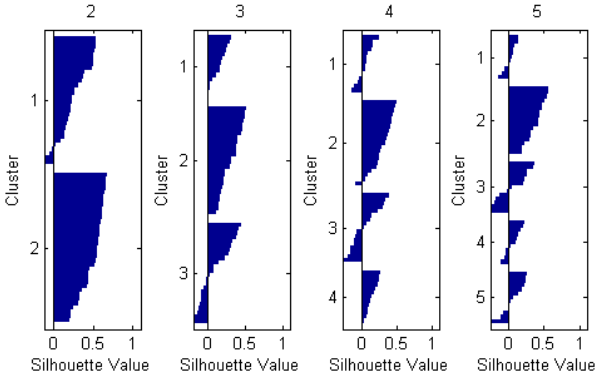


FIGURE 6.7: Silhouette values for different clusters for mouse class I MHC H2-Db allele.

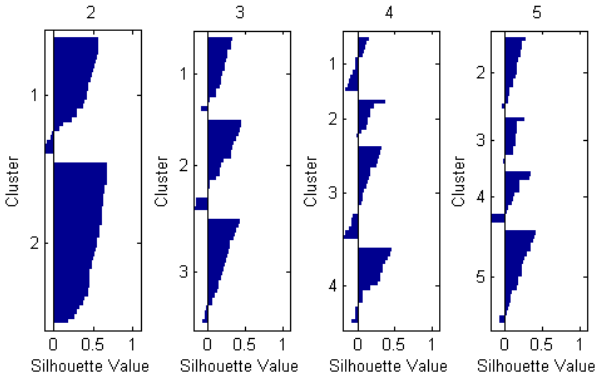


FIGURE 6.8: Silhouette values for different clusters for mouse class I MHC H2-Kb allele.

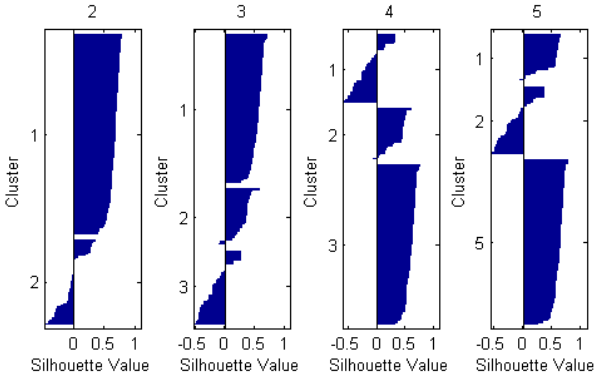


FIGURE 6.9: Silhouette values for different clusters for mouse class I MHC H2-Kk allele.

TABLE 6.1: The optimal TSK-SVR II model parameter values for each peptide binding affinity data set (Tasks 1-4) with different clustering methods.

Task 1

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	161	0.65	0.05	1.03	-0.01
fuzzy c-means	2	161	0.65	0.05	1.03	-0.01
hierarchical	2	161	0.65	0.05	1.03	-0.01

Task 2

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	247	1.90	0.10	1.45	0.03
fuzzy c-means	2	247	1.90	0.10	1.45	0.03
hierarchical	2	250	1.55	0.10	1.00	-0.03

Task 3

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	172	1.45	0.90	0.83	0.07
fuzzy c-means	2	172	1.45	0.90	0.83	0.06
hierarchical	2	172	1.45	0.90	0.83	0.06

Task 4

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	141	2.30	0.45	1.00	0.10
fuzzy c-means	2	141	2.30	0.45	1.00	0.10
hierarchical	2	141	2.30	0.45	1.00	0.10

TABLE 6.2: The optimal (q^2) TSK-SVR II model parameter values for each mouse class I allele leave-one-out cross validated prediction with different clustering methods.

H2Db

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	37	0.45	0.05	0.99	-0.04
fuzzy c-means	2	36	0.75	0.10	0.99	-0.02
hierarchical	2	36	0.75	0.10	0.98	0.01

H2Kb

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	32	1.40	0.45	1.01	0.00
fuzzy c-means	2	34	1.00	0.05	0.96	0.08
hierarchical	2	25	1.75	0.45	1.00	-0.06

H2Kk

clustering method	number of clusters	number of selected features	C	ϵ	q	p
k-means	2	22	6.95	0.35	1.00	0.01
fuzzy c-means	2	20	4.75	0.20	1.00	0.01
hierarchical	2	18	1.50	0.05	0.96	0.01

6.3 Results and Discussion

A hybrid learning system that incorporates the Type-2 TSK fuzzy system with SVR and clustering methods is proposed and applied to the real value prediction of peptide binding affinity. The consequent parameters of the fuzzy system obtained by SVR whereas antecedent parameters obtained using clustering methods. In order to address computational cost of a type-reduction process, our approach used a different inference engine in which type-reduction is not necessary. To initialize the parameters of IT2 fuzzy sets, a novel clustering concept is developed. This clustering approach is based on the overlapping concept. The experimental results and findings of the proposed method are validated using peptide binding affinity data sets that are different and independent from each other. Two groups of peptide binding affinity data sets in this research study are used for the performance evaluation and verification of the proposed method. The first group of data sets consists of CoEPrA data sets. A blind validation performed to evaluate the performance of the proposed method by using these data sets. The second group of data sets consists of mouse class I MHC alleles. These data sets are used for the evaluation of the performance of our method using cross-validation.

6.3.1 Blind-Validated Peptide Binding Affinity Prediction

The analysis of the peptides that have numeric degree of peptide binding is essential for the design of a model that can predict the quantitative binding affinities of unseen peptides. The performance indicates how good the model finds an accurate binding affinity relationship between peptide and a protein.

The optimal set of parameter values found at the end of grid-search for the peptide binding affinity data sets are used. These parameter values are obtained at the end of the intensive seeking process for each of the experimental data set. Suprisingly, they are similar to the parameters of the respective TSK-SVR I models found in the previous chapter. Nevertheless, TSK-SVR II model requires two more additional parameters (q and p) to be optimised. These parameters come from the defuzzification stage of the model. The optimal parameter values of SVR remained same in seeking for the design parameter values of the defuzzification stage of TSK-SVR II. The use of similar parameter values not only aid for getting the benefit from the findings of TSK-SVR I but

also yield a comparative assessment of these parameter values on the results between TSK-SVR I and TSK-SVR II models. The optimal parameters of linear kernel SVR (C and ϵ) and number of selected features along with the optimal design parameters of BMM method that weights the lower and upper firing strengths that resulted in best performance are given in Table 6.1.

For some tasks, selected features are very similar with the features found in the previous chapter. Therefore, the analysis provided in the previous chapter for the feature selection is almost repeated for the SVR based interval type-2 fuzzy models. Only the set of features from the best model selected for each task is taken into consideration. The amino acid features that contributed most to the efficiency of the proposed models are given in Table 6.3 - Table 6.6. For Task 1, eight amino acid features contributed to the output in more than four separate locations. The amino acid feature numbered with 481 (Hydrophobicity coefficient in reversed phase high performance liquid chromatography) contributed highest as it is represented in seven separate locations on each of the nona-peptide within the data set. This finding suggests that hydrophobic effect is important in mediating the binding process between the peptide and MHC molecule in this data set. Therefore, peptides can be shielded from the surrounding solvent and can be buried inner side of the protein [303]. For Task 2, eleven amino acid features contributed to the output in more than five separate locations. The amino acid feature numbered with 364 (Zimm-Bragg parameter sigma x 1.0E4) contributed highest as it is represented in seven separate locations on each of the octa-peptide within the data set. This finding suggests that helix formation in peptides is important in mediating the binding process between the peptide and MHC molecule in this data set. One main reason for the peptides that can nucleate a helix formation is that the ability of their side chains to participate in hydrophobic bonding [304]. For Task 3, nineteen amino acid features contributed to the output in more than three separate locations. The amino acid features numbered with 110 (Composition), 338 (Relative preference value at C"), 376 (Relative population of conformational state A), 405 (Normalized positional residue frequency at helix termini N") contributed highest as they are represented in four separate locations on each of the nona-peptide within the data set. For Task 4, ten amino acid features contributed to the output in more than three separate locations. The amino acid features numbered with 306 (Average relative fractional occurrence in A0(i-1)), 338 (Relative preference value at C"), 110 (Composition), 125 (Normalized relative frequency of double bend)

contributed highest as they are represented in seven separate locations on each of the nona-peptide within the data set. The amino acid feature numbered with 400 (Polarity) appeared in Task 1, Task 2 and Task 3 as a common feature with location occurrences of 4, 6 and 3, respectively. Therefore, the polarity of an amino acid is considered as one of the highly discriminating feature in these data sets. This finding suggests that polarity is important in mediating the binding process between the peptide and MHC molecule in this data set. It is reported that polarity of amino acids can play important role for the protein ubiquitination process. [305]. The full descriptions of amino acid features can be found in Appendix A.

Table 6.7 depicts prediction results of the peptide binding affinity tasks. As the model parameters are similar to the TSK-SVR I fuzzy system, the comparison between the models is more consistent. IT2 fuzzy system has close but better results than its type-1 counterpart for all tasks except Task 1. For Task 2, IT2 fuzzy system initialized with hierarchical clustering method outperformed the IT2 fuzzy systems initialized with partitional clustering methods. For Tasks 3 and 4, IT2 fuzzy system initialized with any clustering method yielded the same result exactly. The outcomes of the experiments with different clustering methods clearly highlights the initialization strength of overlapping clustering concept for interval type-2 fuzzy systems. The results also suggest that SVR concept has positively contributed the learning of the consequent parameters of IT2 fuzzy system. For any clustering method used for the initialization of IT2 fuzzy system, the results seems identical or very close to each other. Table 6.8 - Table 6.11 depict the improvement gain or loss of peptide binding affinity tasks achieved by the proposed models with respect to each other. It is believed that as the improvement gain or loss of different clustering methods are very close to each other and the optimal parameter values are obtained at the end of intensive searching process, IT2 fuzzy system models become saturated. Even then, for Tasks 2, 3 and 4, SVR based IT2-TSK fuzzy system slightly outperformed the type-1 TSK fuzzy model. For Task 1, no improvement gain for SVR based IT2-TSK fuzzy system is obtained over the type-1 TSK fuzzy model. Overall improvement gain or loss value of TSK-SVR II that uses for k-means, fuzzy c-means and hierarchical clustering methods with respect to TSK-SVR I is 0.56, 0.56 and 1.07, respectively. Interestingly, these values are very close to each other. We believe that this is due to the fact that the partitions obtained for all the clustering methods become saturated.

The proposed method consists of only two rules which allows a simple but a robust FS rule-base [5]. Moreover, the model suggested better results than what has been presented in recently published papers [81], [285], [291], [292]. The results also appear to suggest that different clustering methods other than mentioned in this thesis can also be used for the overlapping concept. Further exploration of clustering methods in overlapping concept may improve the initialization performance of antecedent parameters of IT2 fuzzy systems.

TABLE 6.3: Top most frequent amino acid features selected for the optimal model of Task 1 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	481	7	1	1	1	0	1	0	1	1	1
2	302	6	0	1	1	0	1	1	1	0	1
3	367	6	1	1	0	0	1	1	0	1	1
4	31	5	0	0	1	1	0	1	1	1	0
5	613	5	1	1	0	0	0	1	1	0	1
6	259	4	0	1	0	1	0	1	0	1	0
7	359	4	0	0	1	1	0	0	1	1	0
8	400	4	0	1	0	1	0	0	0	1	1

TABLE 6.4: Top most frequent amino acid features selected for the optimal model of Task 2 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	364	7	1	1	0	1	1	1	1	1
2	31	6	1	1	1	1	1	0	0	1
3	379	6	1	0	0	1	1	1	1	1
4	400	6	1	1	0	1	0	1	1	1
5	476	6	1	0	0	1	1	1	1	1
6	30	5	1	0	1	1	0	0	1	1
7	235	5	0	1	1	1	1	0	1	0
8	302	5	0	1	1	1	0	0	1	1
9	380	5	1	0	0	0	1	1	1	1
10	386	5	0	1	1	1	1	0	1	0
11	609	5	1	1	0	1	1	1	0	0

TABLE 6.5: Top most frequent amino acid features selected for the optimal model of Task 3 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	110	4	0	1	0	1	0	1	0	0	1
2	338	4	0	0	0	1	0	1	1	1	0
3	376	4	0	0	0	1	0	1	1	1	0
4	405	4	1	1	1	0	0	0	1	0	0
5	25	3	0	0	1	1	0	0	0	1	0
6	88	3	0	0	1	1	0	1	0	0	0
7	220	3	0	0	0	1	0	0	1	1	0
8	221	3	1	0	0	0	0	1	0	1	0
9	232	3	0	1	0	1	0	0	0	1	0
10	296	3	1	0	0	1	0	0	0	1	0
11	299	3	0	0	0	0	1	1	0	1	0
12	345	3	0	0	0	0	0	1	1	1	0
13	349	3	0	0	1	0	1	0	0	1	0
14	367	3	1	0	0	0	0	0	1	1	0
15	373	3	1	0	0	0	0	1	0	1	0
16	400	3	1	0	0	0	0	0	1	1	0
17	452	3	1	0	0	1	1	0	0	0	0
18	455	3	0	0	1	1	0	0	0	1	0
19	481	3	0	0	0	0	1	0	1	1	0

TABLE 6.6: Top most frequent amino acid features selected for the optimal model of Task 4 and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	306	4	0	0	0	1	0	1	1	1	0
2	338	4	0	0	0	1	0	1	1	1	0
3	110	3	0	1	0	0	0	1	0	0	1
4	125	3	0	0	0	0	1	1	0	1	0
5	221	3	1	0	0	0	0	1	0	1	0
6	232	3	0	1	0	1	0	0	0	1	0
7	251	3	0	0	0	1	0	0	1	0	1
8	373	3	1	0	0	0	0	1	0	1	0
9	405	3	1	1	1	0	0	0	0	0	0
10	420	3	1	0	0	0	0	1	1	0	0

TABLE 6.7: Prediction results of the peptide binding affinity tasks.

Methods	**TRAINING**				**TESTING**			
	Task 1	Task 2	Task 3	Task 4	Task 1	Task 2	Task 3	Task 4
	q^2	q^2	q^2	ρ	q^2	q^2	q^2	ρ
TSK-SVR I	0.8424	0.9862	0.4986	0.8552	**0.6923**	0.7414	0.3092	0.6376
TSK-SVR II (k-means)	0.8474	0.9862	0.4374	0.8548	0.6921	0.7447	**0.3123**	**0.6428**
TSK-SVR II (fuzzy c-means)	0.8500	0.9862	0.4260	0.8548	0.6921	0.7447	**0.3123**	**0.6428**
TSK-SVR II (hierarchical)	0.8422	0.9843	0.4260	0.8548	0.6921	**0.7599**	**0.3123**	**0.6428**

TABLE 6.8: Improvement achieved by the proposed models with respect to each other for peptide binding affinity Task 1.

TASK 1	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	-0.03%	0%		
Type-2 (FCM)	-0.03%	0%	0%	
Type-2 (HIE)	-0.03%	0%	0%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

TABLE 6.9: Improvement achieved by the proposed models with respect to each other for peptide binding affinity Task 2.

TASK 2	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	0.45%	0%		
Type-2 (FCM)	0.45%	0%	0%	
Type-2 (HIE)	2.50%	2.04%	2.04%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

TABLE 6.10: Improvement achieved by the proposed models with respect to each other for peptide binding affinity Task 3.

TASK 3	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	1.00%	0%		
Type-2 (FCM)	1.00%	0%	0%	
Type-2 (HIE)	1.00%	0%	0%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

TABLE 6.11: Improvement achieved by the proposed models with respect to each other for peptide binding affinity Task 4.

TASK 4	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	0.82%	0%		
Type-2 (FCM)	0.82%	0%	0%	
Type-2 (HIE)	0.82%	0%	0%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

6.3.2 Cross-Validated Peptide Binding Affinity Prediction

For comparison purposes, similar to cases studied in the literature for the prediction of mouse class I MHC alleles, each data set is implemented using the leave-one-out cross validation and evaluated with the cross-validated correlation coefficient. The aim of this comparison is to assess the ability of the proposed approach in predicting binding affinities of unseen peptides. Different than the studies presented in the literature, the entire data set prediction is omitted as it does not provide an independent predictive assessment as compared to the evaluation implementing the LOO-CV. Note that, the purpose is to select a model that can efficiently find the affinity of peptide to a protein.

The correct configuration of the parameters of the predictive models is crucial for their performance. Accordingly, the grid-search method is conducted for each mouse class I MHC allele in a sufficient range. As a result, the models are selected based on the optimal set of parameter values found after an intensive seeking process. These optimal parameter sets resemble the ones found for the TSK-SVR I models. Hence, a more consistent comparative analysis can be made on the results between TSK-SVR I and TSK-SVR II models. BMM method is used at the defuzzification stage of TSK-SVR II. This method has two design parameters (q and p) required to be set. They are searched within a sufficient range while the parameters of SVR remained constant. The optimal values of SVR linear kernel parameters and design parameters of BMM method along with the number of selected features yielded best models are given in Table 6.2.

For some alleles, selected features are very similar to the ones that are found in the previous chapter. For this reason, the analysis provided for the feature selection is almost repeated for the SVR based interval type-2 fuzzy models. The amino acid features that contributed most to the efficiency of the proposed models are given in Table 6.12 - Table 6.14. For H2-Db, five amino acid features contributed to the output in two separate locations. The amino acid feature numbered with 18 (Spin-spin coupling constants), 27 (The number of atoms in the side chain), 88 (Positive charge), 481 (Hydrophobicity coefficient in reversed phase high performance liquid chromatography), 520 (Unknown) contributed highest as they are represented in two separate locations on each of the nona-peptide within the data set. For H2-Kb, one amino acid feature contributed to the output in two separate locations. The amino acid feature numbered with 71 (Direction of hydrophobic moment) contributed highest as it is represented in two separate locations

on each of the octa-peptide within the data set. For H2-Kk, three amino acid features contributed to the output in two separate locations. The amino acid feature numbered with 29 (The number of bonds in the longest chain), 88 (Positive charge), 565 (Unknown) contributed highest as they are represented in two separate locations on each of the octa-peptide within the data set. The amino acid feature numbered with 88 (Positive charge) appeared in H2-Db and H2-Kk as a common feature with location occurrences of 2 and 2, respectively. Therefore, the positive charge of an amino acid is considered as one of the highly discriminating feature in these data sets. This finding suggests that positive charge is important in mediating the binding process between the peptide and MHC molecule in this data set. It is reported that positively charged amino acids can play important role for the transmembrane domains of reduced folate carrier [307]. The full descriptions of amino acid features can be found in Appendix A.

Results have been presented in Table 6.15 to ascertain how the proposed models predict the degree of the bindings for the mouse class I MHC alleles. In addition, the results of the TSK-SVR fuzzy system is also provided in order to have consistent comparison among the models. From the table one can see that IT2 fuzzy system models that use the overlapping clustering concept, slightly outperformed the type-1 fuzzy system for the H2-Db, H2-Kb and H2-Kk. On the contrary, for H2-Kk, type-1 fuzzy system yielded better results than IT2 fuzzy system models except for IT2 fuzzy system that uses FCM. Nevertheless, the outcomes of the experiments with different clustering methods clearly highlights the initialization strength of overlapping clustering concept for the interval type-2 fuzzy systems. Interestingly, IT2 fuzzy systems that use FCM achieved slightly better results than those using HCM and hierarchical clustering. The improvement gain or loss of IT2 fuzzy models as well as type-1 fuzzy model in terms of percentages with respect to each other are presented in Table 6.16 - Table 6.18. Overall improvement gain or loss values of TSK-SVR II that uses k-means, fuzzy c-means and hierarchical clustering methods with respect to TSK-SVR I are 1.21, 3.07 and -1.40, respectively. It can be observed that the improvement gain of partitional clustering methods are slightly better than the hierarchical clustering method. A possible reason for this is that the final prototype values of partitional clustering are sensitive to randomly initialized prototype values.

To summarize, the results suggest that SVR concept and overlapping concept have positively contributed the learning of the premise and consequent parameters of IT2 fuzzy

system. For any clustering method used for the initialization of the IT2 fuzzy system, the results seemed identical or very close to each other. Moreover, the proposed models are formed of only two rules which yielded simple and interpretable fuzzy system rule-base [5]. It should also be noted that the models suggested better results than those presented in recently published papers [25], [286], [306]. It can be considered from the results that clustering methods other than those mentioned in this thesis can be incorporated with the overlapping concept. Further exploration of clustering methods in overlapping concept may improve the initialization performance of antecedent parameters of IT2 fuzzy systems.

TABLE 6.12: Top most frequent amino acid features selected for the optimal model of H2-Db and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	18	2	0	0	0	1	0	0	1	0	0
2	27	2	0	0	1	0	0	0	0	0	1
3	88	2	0	1	0	1	0	0	0	0	0
4	481	2	0	0	0	0	0	0	1	0	1

TABLE 6.13: Top most frequent amino acid features selected for the optimal model of H2-Kb and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	71	2	0	1	0	0	1	0	0	0

TABLE 6.14: Top most frequent amino acid features selected for the optimal model of H2-Kk and their appearances on peptide locations.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	29	2	0	1	0	0	0	0	0	1
2	88	2	1	0	0	0	0	0	0	1
3	565	2	0	0	0	0	1	0	1	0

TABLE 6.15: Leave-one-out cross validated correlation coefficient (q^2) prediction results of the mouse class I MHC alleles.

	Allele		
	$H2-D^b$	$H2-K^b$	$H2-K^k$
Methods	q^2	q^2	q^2
TSK-SVR I	0.4624	0.4904	0.7287
TSK-SVR II (k-means)	0.4643	0.5091	0.7245
TSK-SVR II (fuzzy c-means)	**0.4644**	**0.5179**	**0.7519**
TSK-SVR II (hierarchical)	0.4642	0.4920	0.6928

TABLE 6.16: Improvement achieved by the proposed models with respect to each other for H2-Db allele.

H2-Db	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	0.41%	0%		
Type-2 (FCM)	0.43%	0.02%	0%	
Type-2 (HIE)	0.39%	-0.02%	-0.04%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

TABLE 6.17: Improvement achieved by the proposed models with respect to each other for H2-Kb allele.

H2-Kb	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	3.81%	0%		
Type-2 (FCM)	5.61%	1.73%	0%	
Type-2 (HIE)	0.33%	-3.36%	-5.00%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

TABLE 6.18: Improvement achieved by the proposed models with respect to each other for H2-Kk allele.

H2-Kk	Type-1	Type-2 (HCM)	Type-2 (FCM)	Type-2 (HIE)
Type-1	0%			
Type-2 (HCM)	-0.58%	0%		
Type-2 (FCM)	3.18%	3.78%	0%	
Type-2 (HIE)	-4.93%	-4.38%	-7.86%	0%

(HCM) Hard c-Means
(FCM) Fuzzy c-Means
(HIE) HIErarchical

6.4 Conclusions

This chapter has presented a hybrid system and yielded a substantial improvement in the predictive capability of FS with the aid of SVR. During the fuzzification stage, IT2 fuzzy sets are initialized using a novel approach based on overlapping clustering concept. The proposed method was applied to prediction of peptide binding affinity which is regarded as one of the challenging modelling problems in bioinformatics area. Four different peptide binding affinity data sets and three mouse class I MHC alleles were used in order to carry out the experiments. The proposed hybrid system yielded improvements in results than recently published papers that used the same data sets. The prediction results for the proposed method also showed that Type-2 FS has helped to minimise the affects of uncertainties that may exist in the peptide binding affinity data sets and improved the results as compared to its Type-1 counterpart. Apart from improving the prediction accuracy, this research study has also identified amino acid features "Polarity", "Positive charge", "Hydrophobicity coefficient", and "Zimm-Bragg parameter" being the highly discriminating features in the peptide binding affinity data sets.

Chapter 7

Discussion and Conclusions

This thesis has been concentrated to cover research on modelling non-linear systems in the post-genome era. Regression and clustering methods are used to propose a rule-based fuzzy system for the quantitative prediction and analysis of the problems in application domains of bioinformatics. Therefore, combined areas in relation to research, namely fuzzy logic, clustering, regression and feature selection were reviewed in this research study to effectively address modelling non-linear systems for post-genome data sets. This research study introduced two novel methods. First, support-vector based regression is used to identify the structure and parameter values of the consequent part in fuzzy modelling using a closed mathematical form. Second, overlapping clustering concept is used to derive the interval type-2 parameters of the premise part in type-2 fuzzy modelling. Apart from improving the prediction accuracy, this research study has also identified specific features which play a key role(s) in making reliable peptide binding affinity predictions.

This chapter draws conclusions to end the thesis. Summary of the research study is presented in Section 7.1. Strengths and weaknesses of this research study are provided in Section 7.2. Research contributions to literature are provided in Section 7.3. Discussions for further and future work are given in Section 7.4.

7.1 Summary of the Research Study

Fuzzy systems are one of the computational methods which are commonly used to minimise and model uncertainties in the form of rule-based fuzzy logic systems. One advantage of fuzzy systems is that their rule set consists of interpretable IF-THEN rules. There are also some disadvantages when using the fuzzy logic systems. One of the disadvantages of fuzzy systems are coming from their lack of learning capabilities. To increase the learning capabilities of fuzzy systems, a common approach is to combine them with neural networks (e.g. neuro fuzzy systems) or genetic algorithms (e.g. genetic fuzzy systems). Nevertheless, with the increase in size of parameters, the neuro fuzzy systems may become inefficient and a problem what so called curse of dimensionality can be occurred. One aim of this thesis is to develop a novel method and investigate possible solutions to overcome this drawback in fuzzy systems. One possible solution to this problem is to use support vector machines which is a computational method that has a wide use in bioinformatics. Support vector based methods are widely used for non-linear systems and provide mechanisms to handle large number of dimensions with a better generalisation ability.

One main issue in construction of fuzzy systems, is forming the rule-base. Fuzzy clustering is one of the well-identified rule generation methods. This thesis aims primarily constructing a complete initial fuzzy model by discovering the number of clusters and partitioning the post-genome data to obtain appropriate parameters of the rule-based fuzzy system. For the structure and parameter identification in type-2 fuzzy modelling, clustering analysis has been performed using clustering methods such as k-means, fuzzy c-Means, and hierarchical clustering.

Chapter 1 is the introduction chapter where the motivation and structure of the thesis is provided. Basic background information related to amino acids, peptides and proteins are also briefly discussed. Peptide binding affinity problem which is the experimental focus of this thesis is also introduced in this chapter.

Chapter 2 reviews the use of quantitative methods in bioinformatics. Qualitative predictive models often lack of providing certain and precise knowledge due to the ill-defined classes. Therefore, quantitative predictive models are becoming important in the studies of bioinformatics. This review highlights common real-value prediction problems in

various application domains of bioinformatics and possible solutions proposed for them by means of quantitative methods being most of them regression methods. As there is no such review proposed to our best knowledge, this review will fill a gap for those conducting a research study in bioinformatics or systems biology and need to model their research problems in order to predict the real-values of such problems. In this chapter, a set of applications in various application domains of bioinformatics and systems biology that use feature selection are reviewed. The applications are particularly limited to regression-based models in this research study. These applications often dealt with high-dimensionality and small sample size which are the two main issues in the post-genome era. As presented broadly, a common approach to overcome these issues is the use of feature selection methods.

Chapter 3 presents the background theory of this thesis. Combined areas in relation to this research study, namely fuzzy logic, clustering, regression and feature selection, were extensively studied in this chapter to effectively address modelling non-linear systems for post-genome data sets. The performance measurement metrics for the predictive modelling are provided.

In Chapter 4, before presenting our approach for the quantitative prediction of peptide binding affinity. Peptide data sets and how they encoded into a proper feature space from the provided amino acid indices are broadly discussed. The description of AA indices are given as they contain valuable important insight information on the composition of peptides.

In Chapter 5 the usefulness of SVR-based fuzzy system has been showed with real value prediction of degree of peptide binding which is an important problem of bioinformatics. The improvement in the accuracy of predicting real values clearly demonstrates the performance of the proposed approach. Additionally, specific features are also identified which play a key role(s) in making reliable peptide binding affinity predictions.

Chapter 6 presented an SVR-based interval type-2 fuzzy system that is based on overlapping clustering concept for determining the structure of premise part. A closed form defuzzification method, namely BMM method, is used as a defuzzification process of the fuzzy model. The proposed model dealt with the quantitative prediction of peptide binding affinity. The level of uncertainties in the high-dimensional peptide binding affinity data sets are substantially minimised.

7.2 Strength and Weaknesses

This section presents the strengths and weaknesses of the studies carried out here. In terms of the former, the combination of support vector regression with fuzzy logic improved the generalisation ability of the fuzzy system model. Type-1 fuzzy system and interval type-2 fuzzy system both presented the ability of the proposed model to handle associated uncertainties within the biological data sets. For example, the level of uncertainties in the high-dimensional peptide binding affinity data sets are substantially minimised. The feature selection is conducted as a pre-processing stage for all of the experimental peptide binding affinity case studies. It is observed that the selected features are highly dependent on their data sets. Highly discriminating amino acid descriptors were identified (i.e. *Polarity*, *Positive charge*, *Hydrophobicity coefficient*, and *Zimm-Bragg parameter*) in the feature selection process. For the CoEPrA peptide binding affinity data sets, using approximately 5% of the features was sufficient for finding the optimal results. The polarity of an amino acid was observed as a common feature in most of the peptide data sets. For the mouse class I MHC peptide binding affinity data sets, the features are reduced even more; and approximately 0.5% of the features are adequate for finding the optimal models. The positive charge of an amino acid was observed as a common feature in most of the mouse class I alleles. The results obtained here is promising and presents the feasibility and accuracy of the proposed methods. Compared to the previously published results in the literature, the support vector-based type-1 and support vector-based interval type-2 fuzzy models yield an improvement in the prediction accuracy of the peptide binding affinities.

However it must be noted that, although a higher performance accuracy is achieved, identifying optimal parameters of the proposed model(s) can take longer times in relation to other methods mentioned before. Another issue is the limited availability of peptides and their binding affinities. Even the peptide data set is small in size, its dimensionality is high. This is will cause problems in terms of reliability of the predictions made. Fuzzy systems can also suffer from the 'curse of dimensionality' in high-dimensional systems. Feature selection methods are widely used to address this problem and decrease the dimensionality of the feature space. On the contrary, the feature selection method itself may suffer from the problems of dimensionality when high number of features are selected. This will adversely effect the performance of the predictive model. In

addition, the clustering methods (e.g. hierarchical clustering) themselves that are used in the pre-processing stage for forming the rule-base of the fuzzy system, can be sensitive to dimensionality as the number of dimensions increase. This will effect the reliability of clustering process made.

It can not be claimed that our method is the best solution for every problem in bioinformatics as each individual problem may have different dynamics of its own and alternative methods has been also reported in the literature for quantitative prediction as broadly reviewed in Chapter 2. Moreover, the aim of this thesis is to demonstrate how regression based fuzzy systems do for the given problems in bioinformatics and limited to the binding affinity prediction which is the experimental focus in this thesis. Due to the non-linear, complex and high-dimensional nature of bioinformatics problems, it is no doubt that seeking for better solutions still remains an open research problem.

7.3 Contribution to the Literature

The main results and contributions of this thesis are briefly summarised as follows:

- The overlapping method was developed to determine the initial values of antecedent part of the type-2 fuzzy sets. As far as the literature is concerned, to our best knowledge, the proposed overlapping clustering method seems the first formal clustering based approach that helps determine the values of the parameters of type-2 fuzzy membership functions and set a type-2 fuzzy rule base. This is not only simple but also generalise the clustering-based design of the fuzzy system. (journal article is in preparation [29])

- Prediction of peptide binding affinities are regarded as one of the difficult modelling problems in computational biology. The predicted peptide target values using the proposed fuzzy models with the aid of support vector-based method suggest that the predictive ability and performance are increased. The results evidently highlights the strength of the proposed fuzzy models as they yielded comparatively better results than the presented results in the literature. Moreover, the predictive models can speed up work and cut costs for the identification and evaluation of a novel peptide binding at the wet labs. (conference papers are published [26], [27] and journal article is in preparation [28])

- SVR can be used to obtain the parameters of the consequent part of the IT2-TSK fuzzy system and exhibited a good learning candidate as compared to other combinations including least squares learning. (conference paper is published [30] and journal article is in preparation [29])

- The abilities of IT2-FS to model information and handle uncertainties are better as compared to its counterpart T1-FS. On the contrary, IT2-FS has a computational cost and its processing lasts longer. To address this problem a novel method which integrates the inference engine, namely BMM, with the SVR in the consequent part of the IT2-TSK is developed. The inference engine BMM method has a closed mathematical form and conditions required for the stability of IT2-TSK. (conference paper is published [30] and journal article is in preparation [29])

- A review which highlights common real-value prediction problems in various application domains of bioinformatics and systems biology is proposed and possible solutions in the literature to these bio-problems is presented. Regression based methods and feature selection methods that are used in the proposed models in the literature are thoroughly explored. As there is no such review proposed in the literature to our best knowledge, this review will fill a gap and aid for those conducting a research study in the fields of bioinformatics and systems biology. (journal article is in preparation [31])

- To our best knowledge, for the first time, fuzzy systems are used to reveal the discriminating features that can effect the degree of peptide binding to MHC molecules. The features that is most used in the peptide representation would be very useful and provide insights for drug design and inhibitors. The amino acid features *Polarity*, *Positive charge*, *Hydrophobicity coefficient in reversed phase high performance liquid chromatography*, and *Zimm-Bragg parameter* are considered as highly discriminating features in the peptide binding affinity data sets. This novel finding suggests that one can design peptides having features like these which might involve more biological information when designing drugs and vaccines.

7.4 Future Work

The developments made in this thesis suggests new horizons for a future work. The suggestions for a further and future work are discussed and given below:

- Among different fuzzy systems, there are two models widely used in the literature, namely Mamdani fuzzy system and TSK fuzzy system. This thesis concerned with the TSK fuzzy system. The overlapping clustering concept can also be applied to Mamdani fuzzy systems.

- Fuzzy clustering is one of the main methods used in the structure and parameter identification in fuzzy modelling as discussed in this thesis. There exists different alternatives to the fuzzy modelling using fuzzy clustering suggesting simplicity and efficiency. FCM combined with the Gustafson-Kessel algorithm is one such alternative to identify a collection of fuzzy rules efficiently [27]. The possibilistic c-Means [162] is also a kind of fuzzy clustering method that can be treated in generation of membership functions.

- As seen in the chapter that covers literature review, there are many application domains in bioinformatics and systems biology where the quantitative prediction is used. The models suggested in this thesis are also applicable to other bioinformatics problems. They can be used to improve the performance of bioinformatics problems (e.g. prediction of MHC class II binding peptides) in various application domains.

- BMM method used in this thesis simplifies the defuzzification process of the interval type-2 fuzzy system. There exists more defuzzification methods proposed in the literature. These methods also, if applicable, can be incorporated with the support vector regression in the consequent part of the interval type-2 fuzzy system.

- There are different kinds of feature selection methods. Although Multi-Cluster Feature Selection used in this research study. There are many promising feature selection (e.g. Lasso) can be used as a pre-processing step in the model building process.

- In this thesis, SV-based regression is proposed to be used in the consequent part of the interval type-2 fuzzy system. However, there are many regression methods proposed in the literature (e.g. Ridge Regression, Least Angle Regression etc.). They can also be considered to be used to design the consequent part of an interval type-2 fuzzy system.

Appendix A

Amino Acid Indices

The 643 amino acid indices obtained from CoEPrA modeling competition are used in our experimental studies [285]. In Table A.1, the descriptions of 507 amino acids are provided. These descriptions are discovered from AAindex ver.9.1 [289]. The descriptions of remaining 136 amino acids are unknown. Similar to the columns in AAindex database, columns of Table A.1 contains, if exists AAindex accession number and the description of each index. The supplementary information of this thesis is accessible online at: `https://github.com/vuslan/pepbnd`.

TABLE A.1: Description of Amino Acid Indices

ID	Accession No	Description
1	ANDN920101	alpha-CH chemical shifts
2	ARGP820101	Hydrophobicity index
3	ARGP820102	Signal sequence helical potential
4	ARGP820103	Membrane-buried preference parameters
5	BEGF750101	Conformational parameter of inner helix
6	BEGF750102	Conformational parameter of beta-structure
7	BEGF750103	Conformational parameter of beta-turn
8	BHAR880101	Average flexibility indices
9	BIGC670101	Residue volume
10	BIOV880101	Information value for accessibility; average fraction 35%
11	BIOV880102	Information value for accessibility; average fraction 23%
12	BROC820101	Retention coefficient in TFA
13	BROC820102	Retention coefficient in HFBA
14	BULH740101	Transfer free energy to surface
15	BULH740102	Apparent partial specific volume
16	BUNA790101	alpha-NH chemical shifts
17	BUNA790102	alpha-CH chemical shifts
18	BUNA790103	Spin-spin coupling constants
19	BURA740101	Normalized frequency of alpha-helix
20	BURA740102	Normalized frequency of extended structure
21	CHAM810101	Steric parameter
22	CHAM820101	Polarizability parameter
23	CHAM820102	Free energy of solution in water, kcal/mole
24	CHAM830101	The Chou-Fasman parameter of the coil conformation
25	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet
26	CHAM830103	The number of atoms in the side chain labelled 1+1
27	CHAM830104	The number of atoms in the side chain labelled 2+1
28	CHAM830105	The number of atoms in the side chain labelled 3+1
29	CHAM830106	The number of bonds in the longest chain
30	CHAM830107	A parameter of charge transfer capability
31	CHAM830108	A parameter of charge transfer donor capability
32	CHOC750101	Average volume of buried residue
33	CHOC760101	Residue accessible surface area in tripeptide
34	CHOC760102	Residue accessible surface area in folded protein
35	CHOC760103	Proportion of residues 95% buried
36	CHOC760104	Proportion of residues 100% buried

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
37	CHOP780101	Normalized frequency of beta-turn
38	CHOP780201	Normalized frequency of alpha-helix
39	CHOP780202	Normalized frequency of beta-sheet
40	CHOP780203	Normalized frequency of beta-turn
41	CHOP780204	Normalized frequency of N-terminal helix
42	CHOP780205	Normalized frequency of C-terminal helix
43	CHOP780206	Normalized frequency of N-terminal non helical region
44	CHOP780207	Normalized frequency of C-terminal non helical region
45	CHOP780208	Normalized frequency of N-terminal beta-sheet
46	CHOP780209	Normalized frequency of C-terminal beta-sheet
47	CHOP780210	Normalized frequency of N-terminal non beta region
48	CHOP780211	Normalized frequency of C-terminal non beta region
49	CHOP780212	Frequency of the 1st residue in turn
50	CHOP780213	Frequency of the 2nd residue in turn
51	CHOP780214	Frequency of the 3rd residue in turn
52	CHOP780215	Frequency of the 4th residue in turn
53	CHOP780216	Normalized frequency of the 2nd and 3rd residues in turn
54	CIDH920101	Normalized hydrophobicity scales for alpha-proteins
55	CIDH920102	Normalized hydrophobicity scales for beta-proteins
56	CIDH920103	Normalized hydrophobicity scales for alpha+beta-proteins
57	CIDH920104	Normalized hydrophobicity scales for alpha/beta-proteins
58	CIDH920105	Normalized average hydrophobicity scales
59	COHE430101	Partial specific volume
60	CRAJ730101	Normalized frequency of middle helix
61	CRAJ730102	Normalized frequency of beta-sheet
62	CRAJ730103	Normalized frequency of turn
63	DAWD720101	Size
64	DAYM780101	Amino acid composition
65	DAYM780201	Relative mutability
66	DESM900101	Membrane preference for cytochrome b: MPH89
67	DESM900102	Average membrane preference: AMP07
68	EISD840101	Consensus normalized hydrophobicity scale
69	EISD860101	Solvation free energy
70	EISD860102	Atom-based hydrophobic moment
71	EISD860103	Direction of hydrophobic moment
72	FASG760101	Molecular weight
73	FASG760102	Melting point

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
74	FASG760103	Optical rotation
75	FASG760104	pK-N
76	FASG760105	pK-C
77	FAUJ830101	Hydrophobic parameter pi
78	FAUJ880101	Graph shape index
79	FAUJ880102	Smoothed upsilon steric parameter
80	FAUJ880103	Normalized van der Waals volume
81	FAUJ880104	STERIMOL length of the side chain
82	FAUJ880105	STERIMOL minimum width of the side chain
83	FAUJ880106	STERIMOL maximum width of the side chain
84	FAUJ880107	N.m.r. chemical shift of alpha-carbon
85	FAUJ880108	Localized electrical effect
86	FAUJ880109	Number of hydrogen bond donors
87	FAUJ880110	Number of full nonbonding orbitals
88	FAUJ880111	Positive charge
89	FAUJ880112	Negative charge
90	FAUJ880113	pK-a(RCOOH)
91	FINA770101	Helix-coil equilibrium constant
92	FINA910101	Helix initiation parameter at posision i-1
93	FINA910102	Helix initiation parameter at posision i,i+1,i+2
94	FINA910103	Helix termination parameter at posision j-2,j-1,j
95	FINA910104	Helix termination parameter at posision j+1
96	GARJ730101	Partition coefficient
97	GEIM800101	Alpha-helix indices
98	GEIM800102	Alpha-helix indices for alpha-proteins
99	GEIM800103	Alpha-helix indices for beta-proteins
100	GEIM800104	Alpha-helix indices for alpha/beta-proteins
101	GEIM800105	Beta-strand indices
102	GEIM800106	Beta-strand indices for beta-proteins
103	GEIM800107	Beta-strand indices for alpha/beta-proteins
104	GEIM800108	Aperiodic indices
105	GEIM800109	Aperiodic indices for alpha-proteins
106	GEIM800110	Aperiodic indices for beta-proteins
107	GEIM800111	Aperiodic indices for alpha/beta-proteins
108	GOLD730101	Hydrophobicity factor
109	GOLD730102	Residue volume
110	GRAR740101	Composition

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
111	GRAR740102	Polarity
112	GRAR740103	Volume
113	GUYH850101	Partition energy
114	HOPA770101	Hydration number
115	HOPT810101	Hydrophilicity value
116	HUTJ700101	Heat capacity
117	HUTJ700102	Absolute entropy
118	HUTJ700103	Entropy of formation
119	ISOY800101	Normalized relative frequency of alpha-helix
120	ISOY800102	Normalized relative frequency of extended structure
121	ISOY800103	Normalized relative frequency of bend
122	ISOY800104	Normalized relative frequency of bend R
123	ISOY800105	Normalized relative frequency of bend S
124	ISOY800106	Normalized relative frequency of helix end
125	ISOY800107	Normalized relative frequency of double bend
126	ISOY800108	Normalized relative frequency of coil
127	JANJ780101	Average accessible surface area
128	JANJ780102	Percentage of buried residues
129	JANJ780103	Percentage of exposed residues
130	JANJ790101	Ratio of buried and accessible molar fractions
131	JANJ790102	Transfer free energy
132	JOND750101	Hydrophobicity
133	JOND750102	pK (-COOH)
134	JOND920101	Relative frequency of occurrence
135	JOND920102	Relative mutability
136	JUKT750101	Amino acid distribution
137	JUNJ780101	Sequence frequency
138	KANM800101	Average relative probability of helix
139	KANM800102	Average relative probability of beta-sheet
140	KANM800103	Average relative probability of inner helix
141	KANM800104	Average relative probability of inner beta-sheet
142	KARP850101	Flexibility parameter for no rigid neighbors
143	KARP850102	Flexibility parameter for one rigid neighbor
144	KARP850103	Flexibility parameter for two rigid neighbors
145	KHAG800101	The Kerr-constant increments
146	KLEP840101	Net charge
147	KRIW710101	Side chain interaction parameter

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
148	KRIW790101	Side chain interaction parameter
149	KRIW790102	Fraction of site occupied by water
150	KRIW790103	Side chain volume
151	KYTJ820101	Hydropathy index
152	LAWE840101	Transfer free energy, CHP/water
153	LEVM760101	Hydrophobic parameter
154	LEVM760102	Distance between C-alpha and centroid of side chain
155	LEVM760103	Side chain angle theta(AAR)
156	LEVM760104	Side chain torsion angle phi(AAAR)
157	LEVM760105	Radius of gyration of side chain
158	LEVM760106	van der Waals parameter R0
159	LEVM760107	van der Waals parameter epsilon
160	LEVM780101	Normalized frequency of alpha-helix, with weights
161	LEVM780102	Normalized frequency of beta-sheet, with weights
162	LEVM780103	Normalized frequency of reverse turn, with weights
163	LEVM780104	Normalized frequency of alpha-helix, unweighted
164	LEVM780105	Normalized frequency of beta-sheet, unweighted
165	LEVM780106	Normalized frequency of reverse turn, unweighted
166	LEWP710101	Frequency of occurrence in beta-bends
167	LIFS790101	Conformational preference for all beta-strands
168	LIFS790102	Conformational preference for parallel beta-strands
169	LIFS790103	Conformational preference for antiparallel beta-strands
170	MANP780101	Average surrounding hydrophobicity
171	MAXF760101	Normalized frequency of alpha-helix
172	MAXF760102	Normalized frequency of extended structure
173	MAXF760103	Normalized frequency of zeta R
174	MAXF760104	Normalized frequency of left-handed alpha-helix
175	MAXF760105	Normalized frequency of zeta L
176	MAXF760106	Normalized frequency of alpha region
177	MCMT640101	Refractivity
178	MEEJ800101	Retention coefficient in HPLC, pH7.4
179	MEEJ800102	Retention coefficient in HPLC, pH2.1
180	MEEJ810101	Retention coefficient in NaClO4
181	MEEJ810102	Retention coefficient in NaH2PO4
182	MEIH800101	Average reduced distance for C-alpha
183	MEIH800102	Average reduced distance for side chain
184	MEIH800103	Average side chain orientation angle

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
185	MIYS850101	Effective partition energy
186	NAGK730101	Normalized frequency of alpha-helix
187	NAGK730102	Normalized frequency of bata-structure
188	NAGK730103	Normalized frequency of coil
189	NAKH900101	AA composition of total proteins
190	NAKH900102	SD of AA composition of total proteins
191	NAKH900103	AA composition of mt-proteins
192	NAKH900104	Normalized composition of mt-proteins
193	NAKH900105	AA composition of mt-proteins from animal
194	NAKH900106	Normalized composition from animal
195	NAKH900107	AA composition of mt-proteins from fungi and plant
196	NAKH900108	Normalized composition from fungi and plant
197	NAKH900109	AA composition of membrane proteins
198	NAKH900110	Normalized composition of membrane proteins
199	NAKH900111	Transmembrane regions of non-mt-proteins
200	NAKH900112	Transmembrane regions of mt-proteins
201	NAKH900113	Ratio of average and computed composition
202	NAKH920101	AA composition of CYT of single-spanning proteins
203	NAKH920102	AA composition of CYT2 of single-spanning proteins
204	NAKH920103	AA composition of EXT of single-spanning proteins
205	NAKH920104	AA composition of EXT2 of single-spanning proteins
206	NAKH920105	AA composition of MEM of single-spanning proteins
207	NAKH920106	AA composition of CYT of multi-spanning proteins
208	NAKH920107	AA composition of EXT of multi-spanning proteins
209	NAKH920108	AA composition of MEM of multi-spanning proteins
210	NISK800101	8 A contact number
211	NISK860101	14 A contact number
212	NOZY710101	Transfer energy, organic solvent/water
213	OOBM770101	Average non-bonded energy per atom
214	OOBM770102	Short and medium range non-bonded energy per atom
215	OOBM770103	Long range non-bonded energy per atom
216	OOBM770104	Average non-bonded energy per residue
217	OOBM770105	Short and medium range non-bonded energy per residue
218	OOBM850101	Optimized beta-structure-coil equilibrium constant
219	OOBM850102	Optimized propensity to form reverse turn
220	OOBM850103	Optimized transfer energy parameter
221	OOBM850104	Optimized average non-bonded energy per atom

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
222	OOBM850105	Optimized side chain interaction parameter
223	PALJ810101	Normalized frequency of alpha-helix from LG
224	PALJ810102	Normalized frequency of alpha-helix from CF
225	PALJ810103	Normalized frequency of beta-sheet from LG
226	PALJ810104	Normalized frequency of beta-sheet from CF
227	PALJ810105	Normalized frequency of turn from LG
228	PALJ810106	Normalized frequency of turn from CF
229	PALJ810107	Normalized frequency of alpha-helix in all-alpha class
230	PALJ810108	Normalized frequency of alpha-helix in alpha+beta class
231	PALJ810109	Normalized frequency of alpha-helix in alpha/beta class
232	PALJ810110	Normalized frequency of beta-sheet in all-beta class
233	PALJ810111	Normalized frequency of beta-sheet in alpha+beta class
234	PALJ810112	Normalized frequency of beta-sheet in alpha/beta class
235	PALJ810113	Normalized frequency of turn in all-alpha class
236	PALJ810114	Normalized frequency of turn in all-beta class
237	PALJ810115	Normalized frequency of turn in alpha+beta class
238	PALJ810116	Normalized frequency of turn in alpha/beta class
239	PARJ860101	HPLC parameter
240	PLIV810101	Partition coefficient
241	PONP800101	Surrounding hydrophobicity in folded form
242	PONP800102	Average gain in surrounding hydrophobicity
243	PONP800103	Average gain ratio in surrounding hydrophobicity
244	PONP800104	Surrounding hydrophobicity in alpha-helix
245	PONP800105	Surrounding hydrophobicity in beta-sheet
246	PONP800106	Surrounding hydrophobicity in turn
247	PONP800107	Accessibility reduction ratio
248	PONP800108	Average number of surrounding residues
249	PRAM820101	Intercept in regression analysis
250	PRAM820102	Slope in regression analysis x 1.0E1
251	PRAM820103	Correlation coefficient in regression analysis
252	PRAM900101	Hydrophobicity
253	PRAM900102	Relative frequency in alpha-helix
254	PRAM900103	Relative frequency in beta-sheet
255	PRAM900104	Relative frequency in reverse-turn
256	PTIO830101	Helix-coil equilibrium constant
257	PTIO830102	Beta-coil equilibrium constant
258	QIAN880101	Weights for alpha-helix at the window position of -6

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
259	QIAN880102	Weights for alpha-helix at the window position of -5
260	QIAN880103	Weights for alpha-helix at the window position of -4
261	QIAN880104	Weights for alpha-helix at the window position of -3
262	QIAN880105	Weights for alpha-helix at the window position of -2
263	QIAN880106	Weights for alpha-helix at the window position of -1
264	QIAN880107	Weights for alpha-helix at the window position of 0
265	QIAN880108	Weights for alpha-helix at the window position of 1
266	QIAN880109	Weights for alpha-helix at the window position of 2
267	QIAN880110	Weights for alpha-helix at the window position of 3
268	QIAN880111	Weights for alpha-helix at the window position of 4
269	QIAN880112	Weights for alpha-helix at the window position of 5
270	QIAN880113	Weights for alpha-helix at the window position of 6
271	QIAN880114	Weights for beta-sheet at the window position of -6
272	QIAN880115	Weights for beta-sheet at the window position of -5
273	QIAN880116	Weights for beta-sheet at the window position of -4
274	QIAN880117	Weights for beta-sheet at the window position of -3
275	QIAN880118	Weights for beta-sheet at the window position of -2
276	QIAN880119	Weights for beta-sheet at the window position of -1
277	QIAN880120	Weights for beta-sheet at the window position of 0
278	QIAN880121	Weights for beta-sheet at the window position of 1
279	QIAN880122	Weights for beta-sheet at the window position of 2
280	QIAN880123	Weights for beta-sheet at the window position of 3
281	QIAN880124	Weights for beta-sheet at the window position of 4
282	QIAN880125	Weights for beta-sheet at the window position of 5
283	QIAN880126	Weights for beta-sheet at the window position of 6
284	QIAN880127	Weights for coil at the window position of -6
285	QIAN880128	Weights for coil at the window position of -5
286	QIAN880129	Weights for coil at the window position of -4
287	QIAN880130	Weights for coil at the window position of -3
288	QIAN880131	Weights for coil at the window position of -2
289	QIAN880132	Weights for coil at the window position of -1
290	QIAN880133	Weights for coil at the window position of 0
291	QIAN880134	Weights for coil at the window position of 1
292	QIAN880135	Weights for coil at the window position of 2
293	QIAN880136	Weights for coil at the window position of 3
294	QIAN880137	Weights for coil at the window position of 4
295	QIAN880138	Weights for coil at the window position of 5

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
296	QIAN880139	Weights for coil at the window position of 6
297	RACS770101	Average reduced distance for C-alpha
298	RACS770102	Average reduced distance for side chain
299	RACS770103	Side chain orientational preference
300	RACS820101	Average relative fractional occurrence in A0(i)
301	RACS820102	Average relative fractional occurrence in AR(i)
302	RACS820103	Average relative fractional occurrence in AL(i)
303	RACS820104	Average relative fractional occurrence in EL(i)
304	RACS820105	Average relative fractional occurrence in E0(i)
305	RACS820106	Average relative fractional occurrence in ER(i)
306	RACS820107	Average relative fractional occurrence in A0(i-1)
307	RACS820108	Average relative fractional occurrence in AR(i-1)
308	RACS820109	Average relative fractional occurrence in AL(i-1)
309	RACS820110	Average relative fractional occurrence in EL(i-1)
310	RACS820111	Average relative fractional occurrence in E0(i-1)
311	RACS820112	Average relative fractional occurrence in ER(i-1)
312	RACS820113	Value of theta(i)
313	RACS820114	Value of theta(i-1)
314	RADA880101	Transfer free energy from chx to wat
315	RADA880102	Transfer free energy from oct to wat
316	RADA880103	Transfer free energy from vap to chx
317	RADA880104	Transfer free energy from chx to oct
318	RADA880105	Transfer free energy from vap to oct
319	RADA880106	Accessible surface area
320	RADA880107	Energy transfer from out to in(95%buried)
321	RADA880108	Mean polarity
322	RICJ880101	Relative preference value at N”
323	RICJ880102	Relative preference value at N’
324	RICJ880103	Relative preference value at N-cap
325	RICJ880104	Relative preference value at N1
326	RICJ880105	Relative preference value at N2
327	RICJ880106	Relative preference value at N3
328	RICJ880107	Relative preference value at N4
329	RICJ880108	Relative preference value at N5
330	RICJ880109	Relative preference value at Mid
331	RICJ880110	Relative preference value at C5
332	RICJ880111	Relative preference value at C4

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
333	RICJ880112	Relative preference value at C3
334	RICJ880113	Relative preference value at C2
335	RICJ880114	Relative preference value at C1
336	RICJ880115	Relative preference value at C-cap
337	RICJ880116	Relative preference value at C'
338	RICJ880117	Relative preference value at C"
339	ROBB760101	Information measure for alpha-helix
340	ROBB760102	Information measure for N-terminal helix
341	ROBB760103	Information measure for middle helix
342	ROBB760104	Information measure for C-terminal helix
343	ROBB760105	Information measure for extended
344	ROBB760106	Information measure for pleated-sheet
345	ROBB760107	Information measure for extended without H-bond
346	ROBB760108	Information measure for turn
347	ROBB760109	Information measure for N-terminal turn
348	ROBB760110	Information measure for middle turn
349	ROBB760111	Information measure for C-terminal turn
350	ROBB760112	Information measure for coil
351	ROBB760113	Information measure for loop
352	ROBB790101	Hydration free energy
353	ROSG850101	Mean area buried on transfer
354	ROSG850102	Mean fractional area loss
355	ROSM880101	Side chain hydropathy, uncorrected for solvation
356	ROSM880102	Side chain hydropathy, corrected for solvation
357	ROSM880103	Loss of Side chain hydropathy by helix formation
358	SIMZ760101	Transfer free energy
359	SNEP660101	Principal component I
360	SNEP660102	Principal component II
361	SNEP660103	Principal component III
362	SNEP660104	Principal component IV
363	SUEM840101	Zimm-Bragg parameter s at 20 C
364	SUEM840102	Zimm-Bragg parameter sigma x 1.0E4
365	SWER830101	Optimal matching hydrophobicity
366	TANS770101	Normalized frequency of alpha-helix
367	TANS770102	Normalized frequency of isolated helix
368	TANS770103	Normalized frequency of extended structure
369	TANS770104	Normalized frequency of chain reversal R

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
370	TANS770105	Normalized frequency of chain reversal S
371	TANS770106	Normalized frequency of chain reversal D
372	TANS770107	Normalized frequency of left-handed helix
373	TANS770108	Normalized frequency of zeta R
374	TANS770109	Normalized frequency of coil
375	TANS770110	Normalized frequency of chain reversal
376	VASM830101	Relative population of conformational state A
377	VASM830102	Relative population of conformational state C
378	VASM830103	Relative population of conformational state E
379	VELV850101	Electron-ion interaction potential
380	VENT840101	Bitterness
381	VHEG790101	Transfer free energy to lipophilic phase
382	WARP780101	Average interactions per side chain atom
383	WEBA780101	RF value in high salt chromatography
384	WERD780101	Propensity to be buried inside
385	WERD780102	Free energy change of epsilon(i) to epsilon(ex)
386	WERD780103	Free energy change of alpha(Ri) to alpha(Rh)
387	WERD780104	Free energy change of epsilon(i) to alpha(Rh)
388	WOEC730101	Polar requirement
389	WOLR810101	Hydration potential
390	WOLS870101	Principal property value z1
391	WOLS870102	Principal property value z2
392	WOLS870103	Principal property value z3
393	YUTK870101	Unfolding Gibbs energy in water, pH7.0
394	YUTK870102	Unfolding Gibbs energy in water, pH9.0
395	YUTK870103	Activation Gibbs energy of unfolding, pH7.0
396	YUTK870104	Activation Gibbs energy of unfolding, pH9.0
397	ZASB820101	Dependence of partition coefficient on ionic strength
398	ZIMJ680101	Hydrophobicity
399	ZIMJ680102	Bulkiness
400	ZIMJ680103	Polarity
401	ZIMJ680104	Isoelectric point
402	ZIMJ680105	RF rank
403	AURR980101	Normalized positional residue frequency at helix termini N4’
404	AURR980102	Normalized positional residue frequency at helix termini N”’
405	AURR980103	Normalized positional residue frequency at helix termini N”
406	AURR980104	Normalized positional residue frequency at helix termini N’

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
407	AURR980105	Normalized positional residue frequency at helix termini Nc
408	AURR980106	Normalized positional residue frequency at helix termini N1
409	AURR980107	Normalized positional residue frequency at helix termini N2
410	AURR980108	Normalized positional residue frequency at helix termini N3
411	AURR980109	Normalized positional residue frequency at helix termini N4
412	AURR980110	Normalized positional residue frequency at helix termini N5
413	AURR980111	Normalized positional residue frequency at helix termini C5
414	AURR980112	Normalized positional residue frequency at helix termini C4
415	AURR980113	Normalized positional residue frequency at helix termini C3
416	AURR980114	Normalized positional residue frequency at helix termini C2
417	AURR980115	Normalized positional residue frequency at helix termini C1
418	AURR980116	Normalized positional residue frequency at helix termini Cc
419	AURR980117	Normalized positional residue frequency at helix termini C'
420	AURR980118	Normalized positional residue frequency at helix termini C"
421	AURR980119	Normalized positional residue frequency at helix termini C"'
422	AURR980120	Normalized positional residue frequency at helix termini C4'
423	ONEK900101	Delta G values for the peptides extrapolated to 0 M urea
424	ONEK900102	Helix formation parameters (delta delta G)
425	VINM940101	Normalized flexibility parameters (B-values), average
426	VINM940102	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours
427	VINM940103	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours
428	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours
429	MUNV940101	Free energy in alpha-helical conformation
430	MUNV940102	Free energy in alpha-helical region
431	MUNV940103	Free energy in beta-strand conformation
432	MUNV940104	Free energy in beta-strand region
433	MUNV940105	Free energy in beta-strand region
434	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water
435	KIMC930101	Thermodynamic beta sheet propensity
436	MONM990101	Turn propensity scale for transmembrane helices
437	BLAM930101	Alpha helix propensity of position 44 in T4 lysozyme
438	PARS000101	p-Values of mesophilic proteins based on the distributions of B values
439	PARS000102	p-Values of thermophilic proteins based on the distributions of B values
440	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins
441	KUMS000102	Distribution of amino acid residues in the 18 non-redundant families of mesophilic proteins
442	KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins
443	KUMS000104	Distribution of amino acid residues in the alpha-helices in mesophilic proteins

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
444	TAKK010101	Side-chain contribution to protein stability (kJ/mol)
445	FODM020101	Propensity of amino acids within pi-helices
446	NADH010101	Hydropathy scale based on self-information values in the two-state model (5% accessibility)
447	NADH010102	Hydropathy scale based on self-information values in the two-state model (9% accessibility)
448	NADH010103	Hydropathy scale based on self-information values in the two-state model (16% accessibility)
449	NADH010104	Hydropathy scale based on self-information values in the two-state model (20% accessibility)
450	NADH010105	Hydropathy scale based on self-information values in the two-state model (25% accessibility)
451	NADH010106	Hydropathy scale based on self-information values in the two-state model (36% accessibility)
452	NADH010107	Hydropathy scale based on self-information values in the two-state model (50% accessibility)
453	MONM990201	Averaged turn propensities in a transmembrane helix
454	KOEP990101	Alpha-helix propensity derived from designed sequences
455	KOEP990102	Beta-sheet propensity derived from designed sequences
456	CEDJ970101	Composition of amino acids in extracellular proteins (percent)
457	CEDJ970102	Composition of amino acids in anchored proteins (percent)
458	CEDJ970103	Composition of amino acids in membrane proteins (percent)
459	CEDJ970104	Composition of amino acids in intracellular proteins (percent)
460	CEDJ970105	Composition of amino acids in nuclear proteins (percent)
461	FUKS010101	Surface composition of amino acids in intracellular proteins of thermophiles (percent)
462	FUKS010102	Surface composition of amino acids in intracellular proteins of mesophiles (percent)
463	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent)
464	FUKS010104	Surface composition of amino acids in nuclear proteins (percent)
465	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent)
466	FUKS010106	Interior composition of amino acids in intracellular proteins of mesophiles (percent)
467	FUKS010107	Interior composition of amino acids in extracellular proteins of mesophiles (percent)
468	FUKS010108	Interior composition of amino acids in nuclear proteins (percent)
469	FUKS010109	Entire chain composition of amino acids in intracellular proteins of thermophiles (percent)
470	FUKS010110	Entire chain composition of amino acids in intracellular proteins of mesophiles (percent)
471	FUKS010111	Entire chain composition of amino acids in extracellular proteins of mesophiles (percent)
472	FUKS010112	Entire chain composition of amino acids in nuclear proteins (percent)
473	MITS020101	Amphiphilicity index
474	TSAJ990101	Volumes including the crystallographic waters using the ProtOr
475	TSAJ990102	Volumes not including the crystallographic waters using the ProtOr
476	COSI940101	Electron-ion interaction potential values
477	PONP930101	Hydrophobicity scales
478	WILM950101	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O
479	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O
480	WILM950103	Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O

Continued on next page

Table A.1 – *Continued from previous page*

ID	Accession No	Description
481	WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O
482	KUHL950101	Hydrophilicity scale
483	GUOD860101	Retention coefficient at pH 2
484	JURD980101	Modified Kyte-Doolittle hydrophobicity scale
485	BASU050101	Interactivity scale obtained from the contact matrix
486	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins
487	BASU050103	Interactivity scale obtained by maximizing the mean of correlation coefficient over pairs of sequences sharing the TIM barrel fold
488	SUYM030101	Linker propensity index
489	PUNT030101	Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases
490	PUNT030102	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases
491	GEOR030101	Linker propensity from all dataset
492	GEOR030102	Linker propensity from 1-linker dataset
493	GEOR030103	Linker propensity from 2-linker dataset
494	GEOR030104	Linker propensity from 3-linker dataset
495	GEOR030105	Linker propensity from small dataset (linker length is less than six residues)
496	GEOR030106	Linker propensity from medium dataset (linker length is between six and 14 residues)
497	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues)
498	GEOR030108	Linker propensity from helical (annotated by DSSP) dataset
499	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset
500	ZHOH040101	The stability scale from the knowledge-based atom-atom potential
501	ZHOH040102	The relative stability scale extracted from mutation experiments
502	ZHOH040103	Buriability
503	BAEK050101	Linker index
504	HARY940101	Mean volumes of residues buried in protein interiors
505	PONJ960101	Average volumes of residues
506	DIGM050101	Hydrostatic pressure asymmetry index, PAI
507	BLAS910101	Scaled side chain hydrophobicity values

Appendix B

Amino Acid Scales

The scales of 643 amino acid indices obtained from CoEPrA modeling competition that used in our experimental studies, are given in Table B.1 and Table B.2 [285]. The scales of first 507 AA indices given in Table B.1, are discovered that they are from AAindex ver.9.1 [289]. However, the references of remaining 136 AA indices are unknown. Although, their descriptions are not found in the literature, real-values of remaining 136 scales of AA indices are provided in Table B.2. The supplementary information of this thesis is accessible online at: `https://github.com/vuslan/pepbnd`.

TABLE B.1: Real-values of amino acid indices with known descriptions.

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95	4.17	4.36	4.52	4.66	4.44	4.5	4.35	4.7	4.6	3.95
2	0.61	0.6	0.06	0.46	1.07	0	0.47	0.07	0.61	2.22	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32
3	1.18	0.2	0.23	0.05	1.89	0.72	0.11	0.49	0.31	1.45	3.23	0.06	2.67	1.96	0.76	0.97	0.84	0.77	0.39	1.08
4	1.56	0.45	0.27	0.14	1.23	0.51	0.23	0.62	0.29	1.67	2.93	0.15	2.96	2.03	0.76	0.81	0.91	1.08	0.68	1.14
5	1	0.52	0.35	0.44	0.06	0.44	0.73	0.35	0.6	0.73	1	0.6	1	0.6	0.06	0.35	0.44	0.73	0.44	0.82
6	0.77	0.72	0.55	0.65	0.65	0.72	0.55	0.65	0.83	0.98	0.83	0.55	0.98	0.98	0.55	0.55	0.83	0.77	0.83	0.98
7	0.37	0.84	0.97	0.97	0.84	0.64	0.53	0.97	0.75	0.37	0.53	0.75	0.64	0.53	0.97	0.84	0.75	0.97	0.84	0.37
8	0.357	0.529	0.463	0.511	0.346	0.493	0.497	0.544	0.323	0.462	0.365	0.466	0.295	0.314	0.509	0.507	0.444	0.305	0.42	0.386
9	52.6	109.1	75.7	68.4	68.3	89.7	84.7	36.3	91.9	102	102	105.1	97.7	113.9	73.6	54.9	71.2	135.4	116.2	85.1
10	16	-70	-74	-78	168	-73	-106	-13	50	151	145	-141	124	189	-20	-70	-38	145	53	123
11	44	-68	-72	-91	90	-117	-139	-8	47	100	108	-188	121	148	-36	-60	-54	163	22	117
12	7.3	-3.6	-5.7	-2.9	-9.2	-0.3	-7.1	-1.2	-2.1	6.6	20	-3.7	5.6	19.2	5.1	-4.1	0.8	16.3	5.9	3.5
13	3.9	3.2	-2.8	-2.8	-14.3	1.8	-7.5	-2.3	2	11	15	-2.5	4.1	14.7	5.6	-3.5	1.1	17.8	3.8	2.1
14	-0.2	-0.12	0.08	-0.2	-0.45	0.16	-0.3	0	-0.12	-2.26	-2.46	-0.35	-1.47	-2.33	-0.98	-0.39	-0.52	-2.01	-2.24	-1.56
15	0.691	0.728	0.596	0.558	0.624	0.649	0.632	0.592	0.646	0.809	0.842	0.767	0.709	0.756	0.73	0.594	0.655	0.743	0.743	0.777
16	8.249	8.274	8.747	8.41	8.312	8.411	8.368	8.391	8.415	8.195	8.423	8.408	8.418	8.228	0	8.38	8.236	8.094	8.183	8.436
17	4.349	4.396	4.755	4.765	4.686	4.373	4.295	3.972	4.63	4.224	4.385	4.358	4.513	4.663	4.471	4.498	4.346	4.702	4.604	4.184
18	6.5	6.9	7.5	7	7.7	6	7	5.6	8	7	6.5	6.5	0	9.4	0	6.5	6.9	0	6.8	7
19	0.486	0.262	0.193	0.288	0.2	0.418	0.538	0.12	0.4	0.37	0.42	0.402	0.417	0.318	0.208	0.2	0.272	0.462	0.161	0.379
20	0.288	0.362	0.229	0.271	0.533	0.327	0.262	0.312	0.2	0.411	0.4	0.265	0.375	0.318	0.34	0.354	0.388	0.231	0.429	0.495
21	0.52	0.68	0.76	0.76	0.62	0.68	0.68	0	0.7	1.02	0.98	0.68	0.78	0.7	0.36	0.53	0.5	0.7	0.7	0.76
22	0.046	0.291	0.134	0.105	0.128	0.18	0.151	0	0.23	0.186	0.186	0.219	0.221	0.29	0.131	0.062	0.108	0.409	0.298	0.14
23	-0.368	-1.03	0	2.06	4.53	0.731	1.77	-0.525	0	0.791	1.07	0	0.656	1.06	-2.24	-0.524	0	1.6	4.91	0.401
24	0.71	1.06	1.37	1.21	1.19	0.87	0.84	1.52	1.07	0.66	0.69	0.99	0.59	0.71	1.61	1.34	1.08	0.76	1.07	0.63
25	-0.118	0.124	0.289	0.048	0.083	-0.105	-0.245	0.104	0.138	0.23	-0.052	0.032	-0.258	0.015	0	0.225	0.166	0.158	0.094	0.513
26	0	1	1	1	1	1	1	0	1	2	1	1	1	1	0	1	2	1	1	2
27	0	1	1	1	0	1	1	0	1	1	2	1	1	1	0	0	0	1	1	0
28	0	1	0	0	0	1	1	0	1	0	0	1	1	1	0	0	0	1.5	1	0
29	0	5	2	2	1	3	3	0	3	2	2	4	3	4	0	1	1	5	5	1
30	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
31	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0
32	91.5	202	135.2	124.5	117.7	161.1	155.1	66.4	167.3	168.8	167.9	171.3	170.8	203.4	129.3	99.1	122.1	237.6	203.6	141.7
33	115	225	160	150	135	180	190	75	195	175	170	200	185	210	145	115	140	255	230	155
34	25	90	63	50	19	71	49	23	43	18	23	97	31	24	50	44	47	32	60	18
35	0.38	0.01	0.12	0.15	0.45	0.07	0.18	0.36	0.17	0.6	0.45	0.03	0.4	0.5	0.18	0.22	0.23	0.27	0.15	0.54
36	0.2	0	0.03	0.04	0.22	0.01	0.03	0.18	0.02	0.19	0.16	0	0.11	0.14	0.04	0.08	0.08	0.04	0.03	0.18

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
37	0.66	0.95	1.56	1.46	1.19	0.98	0.74	1.56	0.95	0.47	0.59	1.01	0.6	0.6	1.52	1.43	0.96	0.96	1.14	0.5
38	1.42	0.98	0.67	1.01	0.7	1.11	1.51	0.57	1	1.08	1.21	1.16	1.45	1.13	0.57	0.77	0.83	1.08	0.69	1.06
39	0.83	0.93	0.89	0.54	1.19	1.1	0.37	0.75	0.87	1.6	1.3	0.74	1.05	1.38	0.55	0.75	1.19	1.37	1.47	1.7
40	0.74	1.01	1.46	1.52	0.96	0.96	0.95	1.56	0.95	0.47	0.5	1.19	0.6	0.66	1.56	1.43	0.98	0.6	1.14	0.59
41	1.29	0.44	0.81	2.02	0.66	1.22	2.44	0.76	0.73	0.67	0.58	0.66	0.71	0.61	2.01	0.74	1.08	1.47	0.68	0.61
42	1.2	1.25	0.59	0.61	1.11	1.22	1.24	0.42	1.77	0.98	1.13	1.83	1.57	1.1	0	0.96	0.75	0.4	0.73	1.25
43	0.7	0.34	1.42	0.98	0.65	0.75	1.04	1.41	1.22	0.78	0.85	1.01	0.83	0.93	1.1	1.55	1.09	0.62	0.99	0.75
44	0.52	1.24	1.64	1.06	0.94	0.7	0.59	1.64	1.86	0.87	0.84	1.49	0.52	1.04	1.58	0.93	0.86	0.16	0.96	0.32
45	0.86	0.9	0.66	0.38	0.87	1.65	0.35	0.63	0.54	1.94	1.3	1	1.43	1.5	0.66	0.63	1.17	1.49	1.07	1.69
46	0.75	0.9	1.21	0.85	1.11	0.65	0.55	0.74	0.9	1.35	1.27	0.74	0.95	1.5	0.4	0.79	0.75	1.19	1.96	1.79
47	0.67	0.89	1.86	1.39	1.34	1.09	0.92	1.46	0.78	0.59	0.46	1.09	0.52	0.3	1.58	1.41	1.09	0.48	1.23	0.42
48	0.74	1.05	1.13	1.32	0.53	0.77	0.85	1.68	0.96	0.53	0.59	0.82	0.85	0.44	1.69	1.49	1.16	1.59	1.01	0.59
49	0.06	0.07	0.161	0.147	0.149	0.074	0.056	0.102	0.14	0.043	0.061	0.055	0.068	0.059	0.102	0.12	0.086	0.077	0.082	0.062
50	0.076	0.106	0.083	0.11	0.053	0.098	0.06	0.085	0.047	0.034	0.025	0.115	0.082	0.041	0.301	0.139	0.108	0.013	0.065	0.048
51	0.035	0.099	0.191	0.179	0.117	0.037	0.077	0.19	0.093	0.013	0.036	0.072	0.014	0.065	0.034	0.125	0.065	0.064	0.114	0.028
52	0.058	0.085	0.091	0.081	0.128	0.098	0.064	0.152	0.054	0.056	0.07	0.095	0.055	0.065	0.068	0.106	0.079	0.167	0.125	0.053
53	0.64	1.05	1.56	1.61	0.92	0.84	0.8	1.63	0.77	0.29	0.36	1.13	0.51	0.62	2.04	1.52	0.98	0.48	1.08	0.43
54	-0.45	-0.24	-0.2	-1.52	0.79	-0.99	-0.8	-1	1.07	0.76	1.29	-0.36	1.37	1.48	-0.12	-0.98	-0.7	1.38	1.49	1.26
55	-0.08	-0.09	-0.7	-0.71	0.76	-0.4	-1.31	-0.84	0.43	1.39	1.24	-0.09	1.27	1.53	-0.01	-0.93	-0.59	2.25	1.53	1.09
56	0.36	-0.52	-0.9	-1.09	0.7	-1.05	-0.83	-0.82	0.16	2.17	1.18	-0.56	1.21	1.01	-0.06	-0.6	-1.2	1.31	1.05	1.21
57	0.17	-0.7	-0.9	-1.05	1.24	-1.2	-1.19	-0.57	-0.25	2.06	0.96	-0.62	0.6	1.29	-0.21	-0.83	-0.62	1.51	0.66	1.21
58	0.02	-0.42	-0.77	-1.04	0.77	-1.1	-1.14	-0.8	0.26	1.81	1.14	-0.41	1	1.35	-0.09	-0.97	-0.77	1.71	1.11	1.13
59	0.75	0.7	0.61	0.6	0.61	0.67	0.66	0.64	0.67	0.9	0.9	0.82	0.75	0.77	0.76	0.68	0.7	0.74	0.71	0.86
60	1.33	0.79	0.72	0.97	0.93	1.42	1.66	0.58	1.49	0.99	1.29	1.03	1.4	1.15	0.49	0.83	0.94	1.33	0.49	0.96
61	1	0.74	0.75	0.89	0.99	0.87	0.37	0.56	0.36	1.75	1.53	1.18	1.4	1.26	0.36	0.65	1.15	0.84	1.41	1.61
62	0.6	0.79	1.42	1.24	1.29	0.92	0.64	1.38	0.95	0.67	0.7	1.1	0.67	1.05	1.47	1.26	1.05	1.23	1.35	0.48
63	2.5	7.5	5	2.5	3	6	5	0.5	6	5.5	5.5	7	6	6.5	5.5	3	5	7	7	5
64	8.6	4.9	4.3	5.5	2.9	3.9	6	8.4	2	4.5	7.4	6.6	1.7	3.6	5.2	7	6.1	1.3	3.4	6.6
65	100	65	134	106	20	93	102	49	66	96	40	56	94	41	56	120	97	18	41	74
66	1.56	0.59	0.51	0.23	1.8	0.39	0.19	1.03	1	1.27	1.38	0.15	1.93	1.42	0.27	0.96	1.11	0.91	1.1	1.58
67	1.26	0.38	0.59	0.27	1.6	0.39	0.23	1.08	1	1.44	1.36	0.33	1.52	1.46	0.54	0.98	1.01	1.06	0.89	1.33
68	0.25	-1.76	-0.64	-0.72	0.04	-0.69	-0.62	0.16	-0.4	0.73	0.53	-1.1	0.26	0.61	-0.07	-0.26	-0.18	0.37	0.02	0.54
69	0.67	-2.1	-0.6	-1.2	0.38	-0.22	-0.76	0	0.64	1.9	1.9	-0.57	2.4	2.3	1.2	0.01	0.52	2.6	1.6	1.5
70	0	10	1.3	1.9	0.17	1.9	3	0	0.99	1.2	1	5.7	1.9	1.1	0.18	0.73	1.5	1.6	1.8	0.48
71	0	-0.96	-0.86	-0.98	0.76	-1	-0.89	0	-0.75	0.99	0.89	-0.99	0.94	0.92	0.22	-0.67	0.09	0.67	-0.93	0.84
72	89.09	174.2	132.12	133.1	121.15	146.15	147.13	75.07	155.16	131.17	131.17	146.19	149.21	165.19	115.13	105.09	119.12	204.24	181.19	117.15
73	297	238	236	270	178	185	249	290	277	284	337	224	283	284	222	228	253	282	344	293

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
74	1.8	12.5	-5.6	5.05	-16.5	6.3	12	0	-38.5	12.4	-11	14.6	-10	-34.5	-86.2	-7.5	-28	-33.7	-10	5.63
75	9.69	8.99	8.8	9.6	8.35	9.13	9.67	9.78	9.17	9.68	9.6	9.18	9.21	9.18	10.64	9.21	9.1	9.44	9.11	9.62
76	2.34	1.82	2.02	1.88	1.92	2.17	2.1	2.35	1.82	2.36	2.36	2.16	2.28	2.16	1.95	2.19	2.09	2.43	2.2	2.32
77	0.31	-1.01	-0.6	-0.77	1.54	-0.22	-0.64	0	0.13	1.8	1.7	-0.99	1.23	1.79	0.72	-0.04	0.26	2.25	0.96	1.22
78	1.28	2.34	1.6	1.6	1.77	1.56	1.56	0	2.99	4.19	2.59	1.89	2.35	2.94	2.67	1.31	3.03	3.21	2.94	3.67
79	0.53	0.69	0.58	0.59	0.66	0.71	0.72	0	0.64	0.96	0.92	0.78	0.77	0.71	0	0.55	0.63	0.84	0.71	0.89
80	1	6.13	2.95	2.78	2.43	3.95	3.78	0	4.66	4	4	4.77	4.43	5.89	2.72	1.6	2.6	8.08	6.47	3
81	2.87	7.82	4.58	4.74	4.47	6.11	5.97	2.06	5.23	4.92	4.92	6.89	6.36	4.62	4.11	3.97	4.11	7.68	4.73	4.11
82	1.52	1.52	1.52	1.52	1.52	1.52	1.52	1	1.52	1.9	1.52	1.52	1.52	1.52	1.52	1.52	1.73	1.52	1.52	1.9
83	2.04	6.24	4.37	3.78	3.41	3.53	3.31	1	5.66	3.49	4.45	4.87	4.8	6.02	4.31	2.7	3.17	5.9	6.72	3.17
84	7.3	11.1	8	9.2	14.4	10.6	11.4	0	10.2	16.1	10.1	10.9	10.4	13.9	17.8	13.1	16.7	13.2	13.9	17.2
85	-0.01	0.04	0.06	0.15	0.12	0.05	0.07	0	0.08	-0.01	-0.01	0	0.04	0.03	0	0.11	0.04	0	0.03	0.01
86	0	4	2	1	0	2	1	0	1	0	0	2	0	0	0	1	1	1	1	0
87	0	3	3	4	0	3	4	0	1	0	0	1	0	0	0	2	2	0	2	0
88	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
89	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
90	4.76	4.3	3.64	5.69	3.67	4.54	5.48	3.77	2.84	4.81	4.79	4.27	4.25	4.31	0	3.83	3.87	4.75	4.3	4.86
91	1.08	1.05	0.85	0.85	0.95	0.95	1.15	0.55	1	1.05	1.25	1.15	1.15	1.1	0.71	0.75	0.75	1.1	1.1	0.95
92	1	0.7	1.7	3.2	1	1	1.7	1	1	0.6	1	0.7	1	1	1	1.7	1.7	1	1	0.6
93	1	0.7	1	1.7	1	1	1.7	1.3	1	1	1	0.7	1	1	13	1	1	1	1	1
94	1.2	1.7	1.2	0.7	1	1	0.7	0.8	1.2	0.8	1	1.7	1	1	1	1.5	1	1	1	0.8
95	1	1.7	1	0.7	1	1	0.7	1.5	1	1	1	1.7	1	1	0.1	1	1	1	1	1
96	0.28	0.1	0.25	0.21	0.28	0.35	0.33	0.17	0.21	0.82	1	0.09	0.74	2.18	0.39	0.12	0.21	5.7	1.26	0.6
97	1.29	1	0.81	1.1	0.79	1.07	1.49	0.63	1.33	1.05	1.31	1.33	1.54	1.13	0.63	0.78	0.77	1.18	0.71	0.81
98	1.13	1.09	1.06	0.94	1.32	0.93	1.2	0.83	1.09	1.05	1.13	1.08	1.23	1.01	0.82	1.01	1.17	1.32	0.88	1.13
99	1.55	0.2	1.2	1.55	1.44	1.13	1.67	0.59	1.21	1.27	1.25	1.2	1.37	0.4	0.21	1.01	0.55	1.86	1.08	0.64
100	1.19	1	0.94	1.07	0.95	1.32	1.64	0.6	1.03	1.12	1.18	1.27	1.49	1.02	0.68	0.81	0.85	1.18	0.77	0.74
101	0.84	1.04	0.66	0.59	1.27	1.02	0.57	0.94	0.81	1.29	1.1	0.86	0.88	1.15	0.8	1.05	1.2	1.15	1.39	1.56
102	0.86	1.15	0.6	0.66	0.91	1.11	0.37	0.86	1.07	1.17	1.28	1.01	1.15	1.34	0.61	0.91	1.14	1.13	1.37	1.31
103	0.91	0.99	0.72	0.74	1.12	0.9	0.41	0.91	1.01	1.29	1.23	0.86	0.96	1.26	0.65	0.93	1.05	1.15	1.21	1.58
104	0.91	1	1.64	1.4	0.93	0.94	0.97	1.51	0.9	0.65	0.59	0.82	0.58	0.72	1.66	1.23	1.04	0.67	0.92	0.6
105	0.8	0.96	1.1	1.6	0	1.6	0.4	2	0.96	0.85	0.8	0.94	0.39	1.2	2.1	1.3	0.6	0	1.8	0.8
106	1.1	0.93	1.57	1.41	1.05	0.81	1.4	1.3	0.85	0.67	0.52	0.94	0.69	0.6	1.77	1.13	0.88	0.62	0.41	0.58
107	0.93	1.01	1.36	1.22	0.92	0.83	1.05	1.45	0.96	0.58	0.59	0.91	0.6	0.71	1.67	1.25	1.08	0.68	0.98	0.62
108	0.75	0.75	0.69	0	1	0.59	0	0	0	2.95	2.4	1.5	1.3	2.65	2.6	0	0.45	3	2.85	1.7
109	88.3	181.2	125.1	110.8	112.4	148.7	140.5	60	152.6	168.5	168.5	175.6	162.2	189	122.2	88.7	118.2	227	193	141.4
110	0	0.65	1.33	1.38	2.75	0.89	0.92	0.74	0.58	0	0	0.33	0	0	0.39	1.42	0.71	0.13	0.2	0

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
111	8.1	10.5	11.6	13	5.5	10.5	12.3	9	10.4	5.2	4.9	11.3	5.7	5.2	8	9.2	8.6	5.4	6.2	5.9
112	31	124	56	54	55	85	83	3	96	111	111	119	105	132	32.5	32	61	170	136	84
113	0.1	1.91	0.48	0.78	-1.42	0.95	0.83	0.33	-0.5	-1.13	-1.18	1.4	-1.59	-2.12	0.73	0.52	0.07	-0.51	-0.21	-1.27
114	1	2.3	2.2	6.5	0.1	2.1	6.2	1.1	2.8	0.8	0.8	5.3	0.7	1.4	0.9	1.7	1.5	1.9	2.1	0.9
115	-0.5	3	0.2	3	-1	0.2	3	0	-0.5	-1.8	-1.8	3	-1.3	-2.5	0	0.3	-0.4	-3.4	-2.3	-1.5
116	29.22	26.37	38.3	37.09	50.7	44.02	41.84	23.71	59.64	45	48.03	57.1	69.32	48.52	36.13	32.4	35.2	56.92	51.73	40.35
117	30.88	68.43	41.7	40.66	53.83	46.62	44.98	24.74	65.99	49.71	50.62	63.21	55.32	51.06	39.21	35.65	36.5	60	51.15	42.75
118	154.33	341.01	207.9	194.91	219.79	235.51	223.16	127.9	242.54	233.21	232.3	300.46	202.65	204.74	179.93	174.06	205.8	237.01	229.15	207.6
119	1.53	1.17	0.6	1	0.89	1.27	1.63	0.44	1.03	1.07	1.32	1.26	1.66	1.22	0.25	0.65	0.86	1.05	0.7	0.93
120	0.86	0.98	0.74	0.69	1.39	0.89	0.66	0.7	1.06	1.31	1.01	0.77	1.06	1.16	1.16	1.09	1.24	1.17	1.28	1.4
121	0.78	1.06	1.56	1.5	0.6	0.78	0.97	1.73	0.83	0.4	0.57	1.01	0.3	0.67	1.55	1.19	1.09	0.74	1.14	0.44
122	1.09	0.97	1.14	0.77	0.5	0.83	0.92	1.25	0.67	0.66	0.44	1.25	0.45	0.5	2.96	1.21	1.33	0.62	0.94	0.56
123	0.35	0.75	2.12	2.16	0.5	0.73	0.65	2.4	1.19	0.12	0.58	0.83	0.22	0.89	0.43	1.24	0.85	0.62	1.44	0.43
124	1.09	1.07	0.88	1.24	1.04	1.09	1.14	0.27	1.07	0.97	1.3	1.2	0.55	0.8	1.78	1.2	0.99	1.03	0.69	0.77
125	1.34	2.78	0.92	1.77	1.44	0.79	2.54	0.95	0	0.52	1.05	0.79	0	0.43	0.37	0.87	1.14	1.79	0.73	0
126	0.47	0.52	2.16	1.15	0.41	0.95	0.64	3.03	0.89	0.62	0.53	0.98	0.68	0.61	0.63	1.03	0.39	0.63	0.83	0.76
127	27.8	94.7	60.1	60.6	15.5	68.7	68.2	24.5	50.7	22.8	27.6	103	33.5	25.5	51.5	42	45	34.7	55.2	23.7
128	51	5	22	19	74	16	16	52	34	66	60	3	52	58	25	35	30	49	24	64
129	15	67	49	50	5	56	55	10	34	13	16	85	20	10	45	32	32	17	41	14
130	1.7	0.1	0.4	0.4	4.6	0.3	0.3	1.8	0.8	3.1	2.4	0.05	1.9	2.2	0.6	0.8	0.7	1.6	0.5	2.9
131	0.3	-1.4	-0.5	-0.6	0.9	-0.7	-0.7	0.3	-0.1	0.7	0.5	-1.8	0.4	0.5	-0.3	-0.1	-0.2	0.3	-0.4	0.6
132	0.87	0.85	0.09	0.66	1.52	0	0.67	0.1	0.87	3.15	2.17	1.64	1.67	2.87	2.77	0.07	0.07	3.77	2.67	1.87
133	2.34	1.18	2.02	2.01	1.65	2.17	2.19	2.34	1.82	2.36	2.36	2.18	2.28	1.83	1.99	2.21	2.1	2.38	2.2	2.32
134	0.077	0.051	0.043	0.052	0.02	0.041	0.062	0.074	0.023	0.053	0.091	0.059	0.024	0.04	0.051	0.069	0.059	0.014	0.032	0.066
135	100	83	104	86	44	84	77	50	91	103	54	72	93	51	58	117	107	25	50	98
136	5.3	2.6	3	3.6	1.3	2.4	3.3	4.8	1.4	3.1	4.7	4.1	1.1	2.3	2.5	4.5	3.7	0.8	2.3	4.2
137	685	382	397	400	241	313	427	707	155	394	581	575	132	303	366	593	490	99	292	553
138	1.36	1	0.89	1.04	0.82	1.14	1.48	0.63	1.11	1.08	1.21	1.22	1.45	1.05	0.52	0.74	0.81	0.97	0.79	0.94
139	0.81	0.85	0.62	0.71	1.17	0.98	0.53	0.88	0.92	1.48	1.24	0.77	1.05	1.2	0.61	0.92	1.18	1.18	1.23	1.66
140	1.45	1.15	0.64	0.91	0.7	1.14	1.29	0.53	1.13	1.23	1.56	1.27	1.83	1.2	0.21	0.48	0.77	1.17	0.74	1.1
141	0.75	0.79	0.33	0.31	1.46	0.75	0.46	0.83	0.83	1.87	1.56	0.66	0.86	1.37	0.52	0.82	1.36	0.79	1.08	2
142	1.041	1.038	1.117	1.033	0.96	1.165	1.094	1.142	0.982	1.002	0.967	1.093	0.947	0.93	1.055	1.169	1.073	0.925	0.961	0.982
143	0.946	1.028	1.006	1.089	0.878	1.025	1.036	1.042	0.952	0.892	0.961	1.082	0.862	0.912	1.085	1.048	1.051	0.917	0.93	0.927
144	0.892	0.901	0.93	0.932	0.925	0.885	0.933	0.923	0.894	0.872	0.921	1.057	0.804	0.914	0.932	0.923	0.934	0.803	0.837	0.913
145	49.1	133	-3.6	0	0	20	0	64.6	75.7	18.9	15.6	0	6.8	54.7	43.8	44.4	31	70.5	0	29.5
146	0	1	0	-1	0	0	-1	0	0	0	0	1	0	0	0	0	0	0	0	0
147	4.6	6.5	5.9	5.7	-1	6.1	5.6	7.6	4.5	2.6	3.25	7.9	1.4	3.2	7	5.25	4.8	4	4.35	3.4

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
148	4.32	6.55	6.24	6.04	1.73	6.13	6.17	6.09	5.66	2.31	3.93	7.92	2.44	2.59	7.19	5.37	5.16	2.78	3.58	3.31
149	0.28	0.34	0.31	0.33	0.11	0.39	0.37	0.28	0.23	0.12	0.16	0.59	0.08	0.1	0.46	0.27	0.26	0.15	0.25	0.22
150	27.5	105	58.7	40	44.6	80.7	62	0	79	93.5	93.5	100	94.1	115.5	41.9	29.3	51.3	145.5	117.3	71.5
151	1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2
152	-0.48	-0.06	-0.87	-0.75	-0.32	-0.32	-0.71	0	-0.51	0.81	1.02	-0.09	0.81	1.03	2.03	0.05	-0.35	0.66	1.24	0.56
153	-0.5	3	0.2	2.5	-1	0.2	2.5	0	-0.5	-1.8	-1.8	3	-1.3	-2.5	-1.4	0.3	-0.4	-3.4	-2.3	-1.5
154	0.77	3.72	1.98	1.99	1.38	2.58	2.63	0	2.76	1.83	2.08	2.94	2.34	2.97	1.42	1.28	1.43	3.58	3.36	1.49
155	121.9	121.4	117.5	121.2	113.7	118	118.2	0	118.2	118.9	118.1	122	113.1	118.2	81.9	117.9	117.1	118.4	110	121.7
156	243.2	206.6	207.1	215	209.4	205.4	213.6	300	219.9	217.9	205.6	210.9	204	203.7	237.4	232	226.7	203.7	195.6	220.3
157	0.77	2.38	1.45	1.43	1.22	1.75	1.77	0.58	1.78	1.56	1.54	2.08	1.8	1.9	1.25	1.08	1.24	2.21	2.13	1.29
158	5.2	6	5	5	6.1	6	6	4.2	6	7	7	6	6.8	7.1	6.2	4.9	5	7.6	7.1	6.4
159	0.025	0.2	0.1	0.1	0.1	0.1	0.1	0.025	0.1	0.19	0.19	0.2	0.19	0.39	0.17	0.025	0.1	0.56	0.39	0.15
160	1.29	0.96	0.9	1.04	1.11	1.27	1.44	0.56	1.22	0.97	1.3	1.23	1.47	1.07	0.52	0.82	0.82	0.99	0.72	0.91
161	0.9	0.99	0.76	0.72	0.74	0.8	0.75	0.92	1.08	1.45	1.02	0.77	0.97	1.32	0.64	0.95	1.21	1.14	1.25	1.49
162	0.77	0.88	1.28	1.41	0.81	0.98	0.99	1.64	0.68	0.51	0.58	0.96	0.41	0.59	1.91	1.32	1.04	0.76	1.05	0.47
163	1.32	0.98	0.95	1.03	0.92	1.1	1.44	0.61	1.31	0.93	1.31	1.25	1.39	1.02	0.58	0.76	0.79	0.97	0.73	0.93
164	0.86	0.97	0.73	0.69	1.04	1	0.66	0.89	0.85	1.47	1.04	0.77	0.93	1.21	0.68	1.02	1.27	1.26	1.31	1.43
165	0.79	0.9	1.25	1.47	0.79	0.92	1.02	1.67	0.81	0.5	0.57	0.99	0.51	0.77	1.78	1.3	0.97	0.79	0.93	0.46
166	0.22	0.28	0.42	0.73	0.2	0.26	0.08	0.58	0.14	0.22	0.19	0.27	0.38	0.08	0.46	0.55	0.49	0.43	0.46	0.08
167	0.92	0.93	0.6	0.48	1.16	0.95	0.61	0.61	0.93	1.81	1.3	0.7	1.19	1.25	0.4	0.82	1.12	1.54	1.53	1.81
168	1	0.68	0.54	0.5	0.91	0.28	0.59	0.79	0.38	2.6	1.42	0.59	1.49	1.3	0.35	0.7	0.59	0.89	1.08	2.63
169	0.9	1.02	0.62	0.47	1.24	1.18	0.62	0.56	1.12	1.54	1.26	0.74	1.09	1.23	0.42	0.87	1.3	1.75	1.68	1.53
170	12.97	11.72	11.42	10.85	14.63	11.76	11.89	12.43	12.16	15.67	14.9	11.36	14.39	14	11.37	11.23	11.69	13.93	13.42	15.71
171	1.43	1.18	0.64	0.92	0.94	1.22	1.67	0.46	0.98	1.04	1.36	1.27	1.53	1.19	0.49	0.7	0.78	1.01	0.69	0.98
172	0.86	0.94	0.74	0.72	1.17	0.89	0.62	0.97	1.06	1.24	0.98	0.79	1.08	1.16	1.22	1.04	1.18	1.07	1.25	1.33
173	0.64	0.62	3.14	1.92	0.32	0.8	1.01	0.63	2.05	0.92	0.37	0.89	1.07	0.86	0.5	1.01	0.92	1	1.31	0.87
174	0.17	0.76	2.62	1.08	0.95	0.91	0.28	5.02	0.57	0.26	0.21	1.17	0	0.28	0.12	0.57	0.23	0	0.97	0.24
175	1.13	0.48	1.11	1.18	0.38	0.41	1.02	3.84	0.3	0.4	0.65	1.13	0	0.45	0	0.81	0.71	0.93	0.38	0.48
176	1	1.18	0.87	1.39	1.09	1.13	1.04	0.46	0.71	0.68	1.01	1.05	0.36	0.65	1.95	1.56	1.23	1.1	0.87	0.58
177	4.34	26.66	13.28	12	35.77	17.56	17.26	0	21.81	19.06	18.78	21.29	21.64	29.4	10.93	6.35	11.01	42.53	31.53	13.92
178	0.5	0.8	0.8	-8.2	-6.8	-4.8	-16.9	0	-3.5	13.9	8.8	0.1	4.8	13.2	6.1	1.2	2.7	14.9	6.1	2.7
179	-0.1	-4.5	-1.6	-2.8	-2.2	-2.5	-7.5	-0.5	0.8	11.8	10	-3.2	7.1	13.9	8	-3.7	1.5	18.1	8.2	3.3
180	1.1	-0.4	-4.2	-1.6	7.1	-2.9	0.7	-0.2	-0.7	8.5	11	-1.9	5.4	13.4	4.4	-3.2	-1.7	17.1	7.4	5.9
181	1	-2	-3	-0.5	4.6	-2	1.1	0.2	-2.2	7	9.6	-3	4	12.6	3.1	-2.9	-0.6	15.1	6.7	4.6
182	0.93	0.98	0.98	1.01	0.88	1.02	1.02	1.01	0.89	0.79	0.85	1.05	0.84	0.78	1	1.02	0.99	0.83	0.93	0.81
183	0.94	1.09	1.04	1.08	0.84	1.11	1.12	1.01	0.92	0.76	0.82	1.23	0.83	0.73	1.04	1.04	1.02	0.87	1.03	0.81
184	87	81	70	71	104	66	72	90	90	105	104	65	100	108	78	83	83	94	83	94

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
185	2.36	1.92	1.7	1.67	3.36	1.75	1.74	2.06	2.41	4.17	3.93	1.23	4.22	4.37	1.89	1.81	2.04	3.82	2.91	3.49
186	1.29	0.83	0.77	1	0.94	1.1	1.54	0.72	1.29	0.94	1.23	1.23	1.23	1.23	0.7	0.78	0.87	1.06	0.63	0.97
187	0.96	0.67	0.72	0.9	1.13	1.18	0.33	0.9	0.87	1.54	1.26	0.81	1.29	1.37	0.75	0.77	1.23	1.13	1.07	1.41
188	0.72	1.33	1.38	1.04	1.01	0.81	0.75	1.35	0.76	0.8	0.63	0.84	0.62	0.58	1.43	1.34	1.03	0.87	1.35	0.83
189	7.99	5.86	4.33	5.14	1.81	3.98	6.1	6.91	2.17	5.48	9.16	6.01	2.5	3.83	4.95	6.84	5.77	1.34	3.15	6.65
190	3.73	3.34	2.33	2.23	2.3	2.36	3	3.36	1.55	2.52	3.4	3.36	1.37	1.94	3.18	2.83	2.63	1.15	1.76	2.53
191	5.74	1.92	5.25	2.11	1.03	2.3	2.63	5.66	2.3	9.12	15.36	3.2	5.3	6.51	4.79	7.55	7.51	2.51	4.08	5.12
192	-0.6	-1.18	0.39	-1.36	-0.34	-0.71	-1.16	-0.37	0.08	1.44	1.82	-0.84	2.04	1.38	-0.05	0.25	0.66	1.02	0.53	-0.6
193	5.88	1.54	4.38	1.7	1.11	2.3	2.6	5.29	2.33	8.78	16.52	2.58	6	6.58	5.29	7.68	8.38	2.89	3.51	4.66
194	-0.57	-1.29	0.02	-1.54	-0.3	-0.71	-1.17	-0.48	0.1	1.31	2.16	-1.02	2.55	1.42	0.11	0.3	0.99	1.35	0.2	-0.79
195	5.39	2.81	7.31	3.07	0.86	2.31	2.7	6.52	2.23	9.94	12.64	4.67	3.68	6.34	3.62	7.24	5.44	1.64	5.42	6.18
196	-0.7	-0.91	1.28	-0.93	-0.41	-0.71	-1.13	-0.12	0.04	1.77	1.02	-0.4	0.86	1.29	-0.42	0.14	-0.13	0.26	1.29	-0.19
197	9.25	3.96	3.71	3.89	1.07	3.17	4.8	8.51	1.88	6.47	10.94	3.5	3.14	6.36	4.36	6.26	5.66	2.22	3.28	7.55
198	0.34	-0.57	-0.27	-0.56	-0.32	-0.34	-0.43	0.48	-0.19	0.39	0.52	-0.75	0.47	1.3	-0.19	-0.2	-0.04	0.77	0.07	0.36
199	10.17	1.21	1.36	1.18	1.48	1.57	1.15	8.87	1.07	10.91	16.22	1.04	4.12	9.6	2.24	5.38	5.61	2.67	2.68	11.44
200	6.61	0.41	1.84	0.59	0.83	1.2	1.63	4.88	1.14	12.91	21.66	1.15	7.17	7.76	3.51	6.84	8.89	2.11	2.57	6.3
201	1.61	0.4	0.73	0.75	0.37	0.61	1.5	3.12	0.46	1.61	1.37	0.62	1.59	1.24	0.67	0.68	0.92	1.63	0.67	1.3
202	8.63	6.75	4.18	6.24	1.03	4.76	7.82	6.8	2.7	3.48	8.44	6.25	2.14	2.73	6.28	8.53	4.43	0.8	2.54	5.44
203	10.88	6.01	5.75	6.13	0.69	4.68	9.34	7.72	2.15	1.8	8.03	6.11	3.79	2.93	7.21	7.25	3.51	0.47	1.01	4.57
204	5.15	4.38	4.81	5.75	3.24	4.45	7.05	6.38	2.69	4.4	8.11	5.25	1.6	3.52	5.65	8.04	7.41	1.68	3.42	7
205	5.04	3.73	5.94	5.26	2.2	4.5	6.07	7.09	2.99	4.32	9.88	6.31	1.85	3.72	6.22	8.05	5.2	2.1	3.32	6.19
206	9.9	0.09	0.94	0.35	2.55	0.87	0.08	8.14	0.2	15.25	22.28	0.16	1.85	6.47	2.38	4.17	4.33	2.21	3.42	14.34
207	6.69	6.65	4.49	4.97	1.7	5.39	7.76	6.32	2.11	4.51	8.23	8.36	2.46	3.59	5.2	7.4	5.18	1.06	2.75	5.27
208	5.08	4.75	5.75	5.96	2.95	4.24	6.04	8.2	2.1	4.95	8.03	4.93	2.61	4.36	4.84	6.41	5.87	2.31	4.55	6.07
209	9.36	0.27	2.31	0.94	2.56	1.14	0.94	6.17	0.47	13.73	16.64	0.58	3.93	10.99	1.96	5.58	4.68	2.2	3.13	12.43
210	0.23	-0.26	-0.94	-1.13	1.78	-0.57	-0.75	-0.07	0.11	1.19	1.03	-1.05	0.66	0.48	-0.76	-0.67	-0.36	0.9	0.59	1.24
211	-0.22	-0.93	-2.65	-4.12	4.66	-2.76	-3.64	-1.62	1.28	5.58	5.01	-4.18	3.51	5.27	-3.03	-2.84	-1.2	5.2	2.15	4.45
212	0.5	0	0	0	0	0	0	0	0.5	1.8	1.8	0	1.3	2.5	0	0	0.4	3.4	2.3	1.5
213	-1.895	-1.475	-1.56	-1.518	-2.035	-1.521	-1.535	-1.898	-1.755	-1.951	-1.966	-1.374	-1.963	-1.864	-1.699	-1.753	-1.767	-1.869	-1.686	-1.981
214	-1.404	-0.921	-1.178	-1.162	-1.365	-1.116	-1.163	-1.364	-1.215	-1.189	-1.315	-1.074	-1.303	-1.135	-1.236	-1.297	-1.252	-1.03	-1.03	-1.254
215	-0.491	-0.554	-0.382	-0.356	-0.67	-0.405	-0.371	-0.534	-0.54	-0.762	-0.65	-0.3	-0.659	-0.729	-0.463	-0.455	-0.515	-0.839	-0.656	-0.728
216	-9.48	-16.23	-12.48	-12.14	-12.21	-13.69	-13.82	-7.59	-17.55	-15.61	15.73	-12.37	-15.70	-20.50	-11.89	-10.52	-12.37	-26.17	-20.23	-13.87
217	-7.02	-10.131	-9.424	-9.296	-8.19	-10.044	-10.467	-5.456	-12.15	-9.512	10.52	-9.666	-10.424	-12.485	-8.652	-7.782	-8.764	-14.42	-12.36	-8.778
218	2.01	0.84	0.03	-2.05	1.98	1.02	0.93	0.12	-0.14	3.7	2.73	2.55	1.75	2.68	0.41	1.47	2.39	2.49	2.23	3.5
219	1.34	0.95	2.49	3.32	1.07	1.49	2.2	2.07	1.27	0.66	0.54	0.61	0.7	0.8	2.12	0.94	1.09	-4.65	-0.17	1.32
220	0.46	-1.54	1.31	-0.33	0.2	-1.12	0.48	0.64	-1.31	3.28	0.43	-1.71	0.15	0.52	-0.58	-0.83	-1.52	1.25	-2.21	0.54
221	-2.49	2.55	2.27	8.86	-3.13	1.79	4.04	-0.56	4.22	-10.87	-7.16	-9.97	-4.96	-6.64	5.19	-1.6	-4.75	-17.84	9.25	-3.97

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
222	4.55	5.97	5.56	2.85	-0.78	4.15	5.16	9.14	4.48	2.1	3.24	10.68	2.18	4.37	5.14	6.78	8.6	1.97	2.4	3.81
223	1.3	0.93	0.9	1.02	0.92	1.04	1.43	0.63	1.33	0.87	1.3	1.23	1.32	1.09	0.63	0.78	0.8	1.03	0.71	0.95
224	1.32	1.04	0.74	0.97	0.7	1.25	1.48	0.59	1.06	1.01	1.22	1.13	1.47	1.1	0.57	0.77	0.86	1.02	0.72	1.05
225	0.81	1.03	0.81	0.71	1.12	1.03	0.59	0.94	0.85	1.47	1.03	0.77	0.96	1.13	0.75	1.02	1.19	1.24	1.35	1.44
226	0.9	0.75	0.82	0.75	1.12	0.95	0.44	0.83	0.86	1.59	1.24	0.75	0.94	1.41	0.46	0.7	1.2	1.28	1.45	1.73
227	0.84	0.91	1.48	1.28	0.69	1	0.78	1.76	0.53	0.55	0.49	0.95	0.52	0.88	1.47	1.29	1.05	0.88	1.28	0.51
228	0.65	0.93	1.45	1.47	1.43	0.94	0.75	1.53	0.96	0.57	0.56	0.95	0.71	0.72	1.51	1.46	0.96	0.9	1.12	0.55
229	1.08	0.93	1.05	0.86	1.22	0.95	1.09	0.85	1.02	0.98	1.04	1.01	1.11	0.96	0.91	0.95	1.15	1.17	0.8	1.03
230	1.34	0.91	0.83	1.06	1.27	1.13	1.69	0.47	1.11	0.84	1.39	1.08	0.9	1.02	0.48	1.05	0.74	0.64	0.73	1.18
231	1.15	1.06	0.87	1	1.03	1.43	1.37	0.64	0.95	0.99	1.22	1.2	1.45	0.92	0.72	0.84	0.97	1.11	0.72	0.82
232	0.89	1.06	0.67	0.71	1.04	1.06	0.72	0.87	1.04	1.14	1.02	1	1.41	1.32	0.69	0.86	1.15	1.06	1.35	1.66
233	0.82	0.99	1.27	0.98	0.71	1.01	0.54	0.94	1.26	1.67	0.94	0.73	1.3	1.56	0.69	0.65	0.98	1.25	1.26	1.22
234	0.98	1.03	0.66	0.74	1.01	0.63	0.59	0.9	1.17	1.38	1.05	0.83	0.82	1.23	0.73	0.98	1.2	1.26	1.23	1.62
235	0.69	0	1.52	2.42	0	1.44	0.63	2.64	0.22	0.43	0	1.18	0.88	2.2	1.34	1.43	0.28	0	1.53	0.14
236	0.87	1.3	1.36	1.24	0.83	1.06	0.91	1.69	0.91	0.27	0.67	0.66	0	0.47	1.54	1.08	1.12	1.24	0.54	0.69
237	0.91	0.77	1.32	0.9	0.5	1.06	0.53	1.61	1.08	0.36	0.77	1.27	0.76	0.37	1.62	1.34	0.87	1.1	1.24	0.52
238	0.92	0.9	1.57	1.22	0.62	0.66	0.92	1.61	0.39	0.79	0.5	0.86	0.5	0.96	1.3	1.4	1.11	0.57	1.78	0.5
239	2.1	4.2	7	10	1.4	6	7.8	5.7	2.1	-8	-9.2	5.7	-4.2	-9.2	2.1	6.5	5.2	-10	-1.9	-3.7
240	-2.89	-3.3	-3.41	-3.38	-2.49	-3.15	-2.94	-3.25	-2.84	-1.72	-1.61	-3.31	-1.84	-1.63	-2.5	-3.3	-2.91	-1.75	-2.42	-2.08
241	12.28	11.49	11	10.97	14.93	11.28	11.19	12.01	12.84	14.77	14.1	10.8	14.33	13.43	11.19	11.26	11.65	12.95	13.29	15.07
242	7.62	6.81	6.17	6.18	10.93	6.67	6.38	7.31	7.85	9.99	9.37	5.72	9.83	8.99	6.64	6.93	7.08	8.41	8.53	10.38
243	2.63	2.45	2.27	2.29	3.36	2.45	2.31	2.55	2.57	3.08	2.98	2.12	3.18	3.02	2.46	2.6	2.55	2.85	2.79	3.21
244	13.65	11.28	12.24	10.98	14.49	11.3	12.55	15.36	11.59	14.63	14.01	11.96	13.4	14.08	11.51	11.26	13	12.06	12.64	12.88
245	14.6	13.24	11.79	13.78	15.9	12.02	13.59	14.18	15.35	14.1	16.49	13.28	16.23	14.18	14.1	13.36	14.5	13.9	14.76	16.3
246	10.67	11.05	10.85	10.21	14.15	11.71	11.71	10.95	12.07	12.95	13.07	9.93	15	13.27	10.62	11.18	10.53	11.41	11.52	13.86
247	3.7	2.53	2.12	2.6	3.03	2.7	3.3	3.13	3.57	7.69	5.88	1.79	5.21	6.6	2.12	2.43	2.6	6.25	3.03	7.14
248	6.05	5.7	5.04	4.95	7.86	5.45	5.1	6.16	5.8	7.51	7.37	4.88	6.39	6.62	5.65	5.53	5.81	6.98	6.73	7.62
249	0.305	0.227	0.322	0.335	0.339	0.306	0.282	0.352	0.215	0.278	0.262	0.391	0.28	0.195	0.346	0.326	0.251	0.291	0.293	0.291
250	0.175	0.083	0.09	0.14	0.074	0.093	0.135	0.201	0.125	0.1	0.104	0.058	0.054	0.104	0.136	0.155	0.152	0.092	0.081	0.096
251	0.687	0.59	0.489	0.632	0.263	0.527	0.669	0.67	0.594	0.564	0.541	0.407	0.328	0.577	0.6	0.692	0.713	0.632	0.495	0.529
252	-6.7	51.5	20.1	38.5	-8.4	17.2	34.3	-4.2	12.6	-13	-11.7	36.8	-14.2	-15.5	0.8	-2.5	-5	-7.9	2.9	-10.9
253	1.29	0.96	0.9	1.04	1.11	1.27	1.44	0.56	1.22	0.97	1.3	1.23	1.47	1.07	0.52	0.82	0.82	0.99	0.72	0.91
254	0.9	0.99	0.76	0.72	0.74	0.8	0.75	0.92	1.08	1.45	1.02	0.77	0.97	1.32	0.64	0.95	1.21	1.14	1.25	1.49
255	0.78	0.88	1.28	1.41	0.8	0.97	1	1.64	0.69	0.51	0.59	0.96	0.39	0.58	1.91	1.33	1.03	0.75	1.05	0.47
256	1.1	0.95	0.8	0.65	0.95	1	1	0.6	0.85	1.1	1.25	1	1.15	1.1	0.1	0.75	0.75	1.1	1.1	0.95
257	1	0.7	0.6	0.5	1.9	1	0.7	0.3	0.8	4	2	0.7	1.9	3.1	0.2	0.9	1.7	2.2	2.8	4
258	0.12	0.04	-0.1	0.01	-0.25	-0.03	-0.02	-0.02	-0.06	-0.07	0.05	0.26	0	0.05	-0.19	-0.19	-0.04	-0.06	-0.14	-0.03

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
259	0.26	-0.14	-0.03	0.15	-0.15	-0.13	0.21	-0.37	0.1	-0.03	-0.02	0.12	0	0.12	-0.08	0.01	-0.34	-0.01	-0.29	0.02
260	0.64	-0.1	0.09	0.33	0.03	-0.23	0.51	-0.09	-0.23	-0.22	0.41	-0.17	0.13	-0.03	-0.43	-0.1	-0.07	-0.02	-0.38	-0.01
261	0.29	-0.03	-0.04	0.11	-0.05	0.26	0.28	-0.67	-0.26	0	0.47	-0.19	0.27	0.24	-0.34	-0.17	-0.2	0.25	-0.3	-0.01
262	0.68	-0.22	-0.09	-0.02	-0.15	-0.15	0.44	-0.73	-0.14	-0.08	0.61	0.03	0.39	0.06	-0.76	-0.26	-0.1	0.2	-0.04	0.12
263	0.34	0.22	-0.33	0.06	-0.18	0.01	0.2	-0.88	-0.09	-0.03	0.2	-0.11	0.43	0.15	-0.81	-0.35	-0.37	0.07	-0.31	0.13
264	0.57	0.23	-0.36	-0.46	-0.15	0.15	0.26	-0.71	-0.05	0	0.48	0.16	0.41	0.03	-1.12	-0.47	-0.54	-0.1	-0.35	0.31
265	0.33	0.1	-0.19	-0.44	-0.03	0.19	0.21	-0.46	0.27	-0.33	0.57	0.23	0.79	0.48	-1.86	-0.23	-0.33	0.15	-0.19	0.24
266	0.13	0.08	-0.07	-0.71	-0.09	0.12	0.13	-0.39	0.32	0	0.5	0.37	0.63	0.15	-1.4	-0.28	-0.21	0.02	-0.1	0.17
267	0.31	0.18	-0.1	-0.81	-0.26	0.41	-0.06	-0.42	0.51	-0.15	0.56	0.47	0.58	0.1	-1.33	-0.49	-0.44	0.14	-0.08	-0.01
268	0.21	0.07	-0.04	-0.58	-0.12	0.13	-0.23	-0.15	0.37	0.31	0.7	0.28	0.61	-0.06	-1.03	-0.28	-0.25	0.21	0.16	0
269	0.18	0.21	-0.03	-0.32	-0.29	-0.27	-0.25	-0.4	0.28	-0.03	0.62	0.41	0.21	0.05	-0.84	-0.05	-0.16	0.32	0.11	0.06
270	-0.08	0.05	-0.08	-0.24	-0.25	-0.28	-0.19	-0.1	0.29	-0.01	0.28	0.45	0.11	0	-0.42	0.07	-0.33	0.36	0	-0.13
271	-0.18	-0.13	0.28	0.05	-0.26	0.21	-0.06	0.23	0.24	-0.42	-0.23	0.03	-0.42	-0.18	-0.13	0.41	0.33	-0.1	-0.1	-0.07
272	-0.01	0.02	0.41	-0.09	-0.27	0.01	0.09	0.13	0.22	-0.27	-0.25	0.08	-0.57	-0.12	0.26	0.44	0.35	-0.15	0.15	-0.09
273	-0.19	0.03	0.02	-0.06	-0.29	0.02	-0.1	0.19	-0.16	-0.08	-0.42	-0.09	-0.38	-0.32	0.05	0.25	0.22	-0.19	0.05	-0.15
274	-0.14	0.14	-0.27	-0.1	-0.64	-0.11	-0.39	0.46	-0.04	0.16	-0.57	0.04	0.24	0.08	0.02	-0.12	0	-0.1	0.18	0.29
275	-0.31	0.25	-0.53	-0.54	-0.06	0.07	-0.52	0.37	-0.32	0.57	0.09	-0.29	0.29	0.24	-0.31	0.11	0.03	0.15	0.29	0.48
276	-0.1	0.19	-0.89	-0.89	0.13	-0.04	-0.34	-0.45	-0.34	0.95	0.32	-0.46	0.43	0.36	-0.91	-0.12	0.49	0.34	0.42	0.76
277	-0.25	-0.02	-0.77	-1.01	0.13	-0.12	-0.62	-0.72	-0.16	1.1	0.23	-0.59	0.32	0.48	-1.24	-0.31	0.17	0.45	0.77	0.69
278	-0.26	-0.09	-0.34	-0.55	0.47	-0.33	-0.75	-0.56	-0.04	0.94	0.25	-0.55	-0.05	0.2	-1.28	-0.28	0.08	0.22	0.53	0.67
279	0.05	-0.11	-0.4	-0.11	0.36	-0.67	-0.35	0.14	0.02	0.47	0.32	-0.51	-0.1	0.2	-0.79	0.03	-0.15	0.09	0.34	0.58
280	-0.44	-0.13	0.05	-0.2	0.13	-0.58	-0.28	0.08	0.09	-0.04	-0.12	-0.33	-0.21	-0.13	-0.48	0.27	0.47	-0.22	-0.11	0.06
281	-0.31	-0.1	0.06	0.13	-0.11	-0.47	-0.05	0.45	-0.06	-0.25	-0.44	-0.44	-0.28	-0.04	-0.29	0.34	0.27	-0.08	0.06	0.11
282	-0.02	0.04	0.03	0.11	-0.02	-0.17	0.1	0.38	-0.09	-0.48	-0.26	-0.39	-0.14	-0.03	-0.04	0.41	0.36	-0.01	-0.08	-0.18
283	-0.06	0.02	0.1	0.24	-0.19	-0.04	-0.04	0.17	0.19	-0.2	-0.46	-0.43	-0.52	-0.33	0.37	0.43	0.5	-0.32	0.35	0
284	-0.05	0.06	0	0.15	0.3	-0.08	-0.02	-0.14	-0.07	0.26	0.04	-0.42	0.25	0.09	0.31	-0.11	-0.06	0.19	0.33	0.04
285	-0.19	0.17	-0.38	0.09	0.41	0.04	-0.2	0.28	-0.19	-0.06	0.34	-0.2	0.45	0.07	0.04	-0.23	-0.02	0.16	0.22	0.05
286	-0.43	0.06	0	-0.31	0.19	0.14	-0.41	-0.21	0.21	0.29	-0.1	0.33	-0.01	0.25	0.28	-0.23	-0.26	0.15	0.09	-0.1
287	-0.19	-0.07	0.17	-0.27	0.42	-0.29	-0.22	0.17	0.17	-0.34	-0.22	0	-0.53	-0.31	0.14	0.22	0.1	-0.15	-0.02	-0.33
288	-0.25	0.12	0.61	0.6	0.18	0.09	-0.12	0.09	0.42	-0.54	-0.55	0.14	-0.47	-0.29	0.89	0.24	0.16	-0.44	-0.19	-0.45
289	-0.27	-0.4	0.71	0.54	0	-0.08	-0.12	1.14	0.18	-0.74	-0.54	0.45	-0.76	-0.47	1.4	0.4	-0.1	-0.46	-0.05	-0.86
290	-0.42	-0.23	0.81	0.95	-0.18	-0.01	-0.09	1.24	0.05	-1.17	-0.69	0.09	-0.86	-0.39	1.77	0.63	0.29	-0.37	-0.41	-1.32
291	-0.24	-0.04	0.45	0.65	-0.38	0.01	0.07	0.85	-0.21	-0.65	-0.8	0.17	-0.71	-0.61	2.27	0.33	0.13	-0.44	-0.49	-0.99
292	-0.14	0.21	0.35	0.66	-0.09	0.11	0.06	0.36	-0.31	-0.51	-0.8	-0.14	-0.56	-0.25	1.59	0.32	0.21	-0.17	-0.35	-0.7
293	0.01	-0.13	-0.11	0.78	-0.31	-0.13	0.09	0.14	-0.56	-0.09	-0.81	-0.43	-0.49	-0.2	1.14	0.13	-0.02	-0.2	0.1	-0.11
294	-0.3	-0.09	-0.12	0.44	0.03	0.24	0.18	-0.12	-0.2	-0.07	-0.18	0.06	-0.44	0.11	0.77	-0.09	-0.27	-0.09	-0.25	-0.06
295	-0.23	-0.2	0.06	0.34	0.19	0.47	0.28	0.14	-0.22	0.42	-0.36	-0.15	-0.19	-0.02	0.78	-0.29	-0.3	-0.18	0.07	0.29

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
296	0.08	-0.01	-0.06	0.04	0.37	0.48	0.36	-0.02	-0.45	0.09	0.24	-0.27	0.16	0.34	0.16	-0.35	-0.04	-0.06	-0.2	0.18
297	0.934	0.962	0.986	0.994	0.9	1.047	0.986	1.015	0.882	0.766	0.825	1.04	0.804	0.773	1.047	1.056	1.008	0.848	0.931	0.825
298	0.941	1.112	1.038	1.071	0.866	1.15	1.1	1.055	0.911	0.742	0.798	1.232	0.781	0.723	1.093	1.082	1.043	0.867	1.05	0.817
299	1.16	1.72	1.97	2.66	0.5	3.87	2.4	1.63	0.86	0.57	0.51	3.9	0.4	0.43	2.04	1.61	1.48	0.75	1.72	0.59
300	0.85	2.02	0.88	1.5	0.9	1.71	1.79	1.54	1.59	0.67	1.03	0.88	1.17	0.85	1.47	1.5	1.96	0.83	1.34	0.89
301	1.58	1.14	0.77	0.98	1.04	1.24	1.49	0.66	0.99	1.09	1.21	1.27	1.41	1	1.46	1.05	0.87	1.23	0.68	0.88
302	0.82	2.6	2.07	2.64	0	0	2.62	1.63	0	2.32	0	2.86	0	0	0	1.23	2.48	0	1.9	1.62
303	0.78	1.75	1.32	1.25	3.14	0.93	0.94	1.13	1.03	1.26	0.91	0.85	0.41	1.07	1.73	1.31	1.57	0.98	1.31	1.11
304	0.88	0.99	1.02	1.16	1.14	0.93	1.01	0.7	1.87	1.61	1.09	0.83	1.71	1.52	0.87	1.14	0.96	1.96	1.68	1.56
305	0.3	0.9	2.73	1.26	0.72	0.97	1.33	3.09	1.33	0.45	0.96	0.71	1.89	1.2	0.83	1.16	0.97	1.58	0.86	0.64
306	0.4	1.2	1.24	1.59	2.98	0.5	1.26	1.89	2.71	1.31	0.57	0.87	0	1.27	0.38	0.92	1.38	1.53	1.79	0.95
307	1.48	1.02	0.99	1.19	0.86	1.42	1.43	0.46	1.27	1.12	1.33	1.36	1.41	1.3	0.25	0.89	0.81	1.27	0.91	0.93
308	0	0	4.14	2.15	0	0	0	6.49	0	0	0	0	0	2.11	1.99	0	1.24	0	1.9	0
309	1.02	1	1.31	1.76	1.05	1.05	0.83	2.39	0.4	0.83	1.06	0.94	1.33	0.41	2.73	1.18	0.77	1.22	1.09	0.88
310	0.93	1.52	0.92	0.6	1.08	0.94	0.73	0.78	1.08	1.74	1.03	1	1.31	1.51	1.37	0.97	1.38	1.12	1.65	1.7
311	0.99	1.19	1.15	1.18	2.32	1.52	1.36	1.4	1.06	0.81	1.26	0.91	1	1.25	0	1.5	1.18	1.33	1.09	1.01
312	17.05	21.25	34.81	19.27	28.84	15.42	20.12	38.14	23.07	16.66	10.89	16.46	20.61	16.26	23.94	19.95	18.92	23.36	26.49	17.06
313	14.53	17.82	13.59	19.78	30.57	22.18	18.19	37.16	22.63	20.28	14.3	14.07	20.61	19.61	52.63	18.56	21.09	19.78	26.36	21.87
314	1.81	-14.92	-6.64	-8.72	1.28	-5.54	-6.81	0.94	-4.66	4.92	4.92	-5.55	2.35	2.98	0	-3.4	-2.57	2.33	-0.14	4.04
315	0.52	-1.32	-0.01	0	0	-0.07	-0.79	0	0.95	2.04	1.76	0.08	1.32	2.09	0	0.04	0.27	2.51	1.63	1.18
316	0.13	-5	-3.04	-2.23	-2.52	-3.84	-3.43	1.45	-5.61	-2.77	-2.64	-3.97	-3.83	-3.74	0	-1.66	-2.31	-8.21	-5.97	-2.05
317	1.29	-13.6	-6.63	0	0	-5.47	-6.02	0.94	-5.61	2.88	3.16	-5.63	1.03	0.89	0	-3.44	-2.84	-0.18	-1.77	2.86
318	1.42	-18.6	-9.67	0	0	-9.31	-9.45	2.39	-11.22	0.11	0.52	-9.6	-2.8	-2.85	0	-5.1	-5.15	-8.39	-7.74	0.81
319	93.7	250.4	146.3	142.6	135.2	177.7	182.9	52.6	188.1	182.2	173.7	215.2	197.6	228.6	0	109.5	142.1	271.6	239.9	157.2
320	-0.29	-2.71	-1.18	-1.02	0	-1.53	-0.9	-0.34	-0.94	0.24	-0.12	-2.05	-0.24	0	0	-0.75	-0.71	-0.59	-1.02	0.09
321	-0.06	-0.84	-0.48	-0.8	1.36	-0.73	-0.77	-0.41	0.49	1.31	1.21	-1.18	1.27	1.27	0	-0.5	-0.27	0.88	0.33	1.09
322	0.7	0.4	1.2	1.4	0.6	1	1	1.6	1.2	0.9	0.9	1	0.3	1.2	0.7	1.6	0.3	1.1	1.9	0.7
323	0.7	0.4	1.2	1.4	0.6	1	1	1.6	1.2	0.9	0.9	1	0.3	1.2	0.7	1.6	0.3	1.1	1.9	0.7
324	0.5	0.4	3.5	2.1	0.6	0.4	0.4	1.8	1.1	0.2	0.2	0.7	0.8	0.2	0.8	2.3	1.6	0.3	0.8	0.1
325	1.2	0.7	0.7	0.8	0.8	0.7	2.2	0.3	0.7	0.9	0.9	0.6	0.3	0.5	2.6	0.7	0.8	2.1	1.8	1.1
326	1.6	0.9	0.7	2.6	1.2	0.8	2	0.9	0.7	0.7	0.3	1	1	0.9	0.5	0.8	0.7	1.7	0.4	0.6
327	1	0.4	0.7	2.2	0.6	1.5	3.3	0.6	0.7	0.4	0.6	0.8	1	0.6	0.4	0.4	1	1.4	1.2	1.1
328	1.1	1.5	0	0.3	1.1	1.3	0.5	0.4	1.5	1.1	2.6	0.8	1.7	1.9	0.1	0.4	0.5	3.1	0.6	1.5
329	1.4	1.2	1.2	0.6	1.6	1.4	0.9	0.6	0.9	0.9	1.1	1.9	1.7	1	0.3	1.1	0.6	1.4	0.2	0.8
330	1.8	1.3	0.9	1	0.7	1.3	0.8	0.5	1	1.2	1.2	1.1	1.5	1.3	0.3	0.6	1	1.5	0.8	1.2
331	1.8	1	0.6	0.7	0	1	1.1	0.5	2.4	1.3	1.2	1.4	2.7	1.9	0.3	0.5	0.5	1.1	1.3	0.4
332	1.3	0.8	0.6	0.5	0.7	0.2	0.7	0.5	1.9	1.6	1.4	1	2.8	2.9	0	0.5	0.6	2.1	0.8	1.4

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
333	0.7	0.8	0.8	0.6	0.2	1.3	1.6	0.1	1.1	1.4	1.9	2.2	1	1.8	0	0.6	0.7	0.4	1.1	1.3
334	1.4	2.1	0.9	0.7	1.2	1.6	1.7	0.2	1.8	0.4	0.8	1.9	1.3	0.3	0.2	1.6	0.9	0.4	0.3	0.7
335	1.1	1	1.2	0.4	1.6	2.1	0.8	0.2	3.4	0.7	0.7	2	1	0.7	0	1.7	1	0	1.2	0.7
336	0.8	0.9	1.6	0.7	0.4	0.9	0.3	3.9	1.3	0.7	0.7	1.3	0.8	0.5	0.7	0.8	0.3	0	0.8	0.2
337	1	1.4	0.9	1.4	0.8	1.4	0.8	1.2	1.2	1.1	0.9	1.2	0.8	0.1	1.9	0.7	0.8	0.4	0.9	0.6
338	0.7	1.1	1.5	1.4	0.4	1.1	0.7	0.6	1	0.7	0.5	1.3	0	1.2	1.5	0.9	2.1	2.7	0.5	1
339	6.5	-0.9	-5.1	0.5	-1.3	1	7.8	-8.6	1.2	0.6	3.2	2.3	5.3	1.6	-7.7	-3.9	-2.6	1.2	-4.5	1.4
340	2.3	-5.2	0.3	7.4	0.8	-0.7	10.3	-5.2	-2.8	-4	-2.1	-4.1	-3.5	-1.1	8.1	-3.5	2.3	-0.9	-3.7	-4.4
341	6.7	0.3	-6.1	-3.1	-4.9	0.6	2.2	-6.8	-1	3.2	5.5	0.5	7.2	2.8	-22.8	-3	-4	4	-4.6	2.5
342	2.3	1.4	-3.3	-4.4	6.1	2.7	2.5	-8.3	5.9	-0.5	0.1	7.3	3.5	1.6	-24.4	-1.9	-3.7	-0.9	-0.6	2.3
343	-2.3	0.4	-4.1	-4.4	4.4	1.2	-5	-4.2	-2.5	6.7	2.3	-3.3	2.3	2.6	-1.8	-1.7	1.3	-1	4	6.8
344	-2.7	0.4	-4.2	-4.4	3.7	0.8	-8.1	-3.9	-3	7.7	3.7	-2.9	3.7	3	-6.6	-2.4	1.7	0.3	3.3	7.1
345	0	1.1	-2	-2.6	5.4	2.4	3.1	-3.4	0.8	-0.1	-3.7	-3.1	-2.1	0.7	7.4	1.3	0	-3.4	4.8	2.7
346	-5	2.1	4.2	3.1	4.4	0.4	-4.7	5.7	-0.3	-4.6	-5.6	1	-4.8	-1.8	2.6	2.6	0.3	3.4	2.9	-6
347	-3.3	0	5.4	3.9	-0.3	-0.4	-1.8	-1.2	3	-0.5	-2.3	-1.2	-4.3	0.8	6.5	1.8	-0.7	-0.8	3.1	-3.5
348	-4.7	2	3.9	1.9	6.2	-2	-4.2	5.7	-2.6	-7	-6.2	2.8	-4.8	-3.7	3.6	2.1	0.6	3.3	3.8	-6.2
349	-3.7	1	-0.6	-0.6	4	3.4	-4.3	5.9	-0.8	-0.5	-2.8	1.3	-1.6	1.6	-6	1.5	1.2	6.5	1.3	-4.6
350	-2.5	-1.2	4.6	0	-4.7	-0.5	-4.4	4.9	1.6	-3.3	-2	-0.8	-4.1	-4.1	5.8	2.5	1.7	1.2	-0.6	-3.5
351	-5.1	2.6	4.7	3.1	3.8	0.2	-5.2	5.6	-0.9	-4.5	-5.4	1	-5.3	-2.4	3.5	3.2	0	2.9	3.2	-6.3
352	-1	0.3	-0.7	-1.2	2.1	-0.1	-0.7	0.3	1.1	4	2	-0.9	1.8	2.8	0.4	-1.2	-0.5	3	2.1	1.4
353	86.6	162.2	103.3	97.8	132.3	119.2	113.9	62.9	155.8	158	164.1	115.5	172.9	194.1	92.9	85.6	106.5	224.6	177.7	141
354	0.74	0.64	0.63	0.62	0.91	0.62	0.62	0.72	0.78	0.88	0.85	0.52	0.85	0.88	0.64	0.66	0.7	0.85	0.76	0.86
355	-0.67	12.1	7.23	8.72	-0.34	6.39	7.35	0	3.82	-3.02	-3.02	6.13	-1.3	-3.24	-1.75	4.35	3.86	-2.86	0.98	-2.18
356	-0.67	3.89	2.27	1.57	-2	2.12	1.78	0	1.09	-3.02	-3.02	2.46	-1.67	-3.24	-1.75	0.1	-0.42	-2.86	0.98	-2.18
357	0.4	0.3	0.9	0.8	0.5	0.7	1.3	0	1	0.4	0.6	0.4	0.3	0.7	0.9	0.4	0.4	0.6	1.2	0.4
358	0.73	0.73	-0.01	0.54	0.7	-0.1	0.55	0	1.1	2.97	2.49	1.5	1.3	2.65	2.6	0.04	0.44	3	2.97	1.69
359	0.239	0.211	0.249	0.171	0.22	0.26	0.187	0.16	0.205	0.273	0.281	0.228	0.253	0.234	0.165	0.236	0.213	0.183	0.193	0.255
360	0.33	-0.176	-0.233	-0.371	0.074	-0.254	-0.409	0.37	-0.078	0.149	0.129	-0.075	-0.092	-0.011	0.37	0.022	0.136	-0.011	-0.138	0.245
361	-0.11	0.079	-0.136	-0.285	-0.184	-0.067	-0.246	-0.073	0.32	0.001	-0.008	0.049	-0.041	0.438	-0.016	-0.153	-0.208	0.493	0.381	-0.155
362	-0.062	-0.167	0.166	-0.079	0.38	-0.025	-0.184	-0.017	0.056	-0.309	-0.264	-0.371	0.077	0.074	-0.036	0.47	0.348	0.05	0.22	-0.212
363	1.071	1.033	0.784	0.68	0.922	0.977	0.97	0.591	0.85	1.14	1.14	0.939	1.2	1.086	0.659	0.76	0.817	1.107	1.02	0.95
364	8	0.1	0.1	70	26	33	6	0.1	0.1	55	33	1	54	18	42	0.1	0.1	77	66	0.1
365	-0.4	-0.59	-0.92	-1.31	0.17	-0.91	-1.22	-0.67	-0.64	1.25	1.22	-0.67	1.02	1.92	-0.49	-0.55	-0.28	0.5	1.67	0.91
366	1.42	1.06	0.71	1.01	0.73	1.02	1.63	0.5	1.2	1.12	1.29	1.24	1.21	1.16	0.65	0.71	0.78	1.05	0.67	0.99
367	0.946	1.128	0.432	1.311	0.481	1.615	0.698	0.36	2.168	1.283	1.192	1.203	0	0.963	2.093	0.523	1.961	1.925	0.802	0.409
368	0.79	1.087	0.832	0.53	1.268	1.038	0.643	0.725	0.864	1.361	1.111	0.735	1.092	1.052	1.249	1.093	1.214	1.114	1.34	1.428
369	1.194	0.795	0.659	1.056	0.678	1.29	0.928	1.015	0.611	0.603	0.595	1.06	0.831	0.377	3.159	1.444	1.172	0.452	0.816	0.64

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
370	0.497	0.677	2.072	1.498	1.348	0.711	0.651	1.848	1.474	0.471	0.656	0.932	0.425	1.348	0.179	1.151	0.749	1.283	1.283	0.654
371	0.937	1.725	1.08	1.64	1.004	1.078	0.679	0.901	1.085	0.178	0.808	1.254	0.886	0.803	0.748	1.145	1.487	0.803	1.227	0.625
372	0.289	1.38	3.169	0.917	1.767	2.372	0.285	4.259	1.061	0.262	0	1.288	0	0.393	0	0.16	0.218	0	0.654	0.167
373	0.328	2.088	1.498	3.379	0	0	0	0.5	1.204	2.078	0.414	0.835	0.982	1.336	0.415	1.089	1.732	1.781	0	0.946
374	0.945	0.364	1.202	1.315	0.932	0.704	1.014	2.355	0.525	0.673	0.758	0.947	1.028	0.622	0.579	1.14	0.863	0.777	0.907	0.561
375	0.842	0.936	1.352	1.366	1.032	0.998	0.758	1.349	1.079	0.459	0.665	1.045	0.668	0.881	1.385	1.257	1.055	0.881	1.101	0.643
376	0.135	0.296	0.196	0.289	0.159	0.236	0.184	0.051	0.223	0.173	0.215	0.17	0.239	0.087	0.151	0.01	0.1	0.166	0.066	0.285
377	0.507	0.459	0.287	0.223	0.592	0.383	0.445	0.39	0.31	0.111	0.619	0.559	0.431	0.077	0.739	0.689	0.785	0.16	0.06	0.356
378	0.159	0.194	0.385	0.283	0.187	0.236	0.206	0.049	0.233	0.581	0.083	0.159	0.198	0.682	0.366	0.15	0.074	0.463	0.737	0.301
379	0.0373	0.0959	0.0036	0.1263	0.0829	0.0761	0.0058	0.0050	0.0242	0	0	0.0371	0.0823	0.0946	0.0198	0.0829	0.0941	0.0548	0.0516	0.0057
380	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1	1
381	-12.04	39.23	4.25	23.22	3.95	2.16	16.81	-7.85	6.28	-18.32	-17.79	9.71	-8.86	-21.98	5.82	-1.54	-4.15	-16.19	-1.51	-16.22
382	10.04	6.18	5.63	5.76	8.89	5.41	5.37	7.99	7.49	8.72	8.79	4.4	9.15	7.98	7.79	7.08	7	8.07	6.9	8.88
383	0.89	0.88	0.89	0.87	0.85	0.82	0.84	0.92	0.83	0.76	0.73	0.97	0.74	0.52	0.82	0.96	0.92	0.2	0.49	0.85
384	0.52	0.49	0.42	0.37	0.83	0.35	0.38	0.41	0.7	0.79	0.77	0.31	0.76	0.87	0.35	0.49	0.38	0.86	0.64	0.72
385	0.16	-0.2	1.03	-0.24	-0.12	-0.55	-0.45	-0.16	-0.18	-0.19	-0.44	-0.12	-0.79	-0.25	-0.59	-0.01	0.05	-0.33	-0.42	-0.46
386	0.15	-0.37	0.69	-0.22	-0.19	-0.06	0.14	0.36	-0.25	0.02	0.06	-0.16	0.11	1.18	0.11	0.13	0.28	-0.12	0.19	-0.08
387	-0.07	-0.4	-0.57	-0.8	0.17	-0.26	-0.63	0.27	-0.49	0.06	-0.17	-0.45	0.03	0.4	-0.47	-0.11	0.09	-0.61	-0.61	-0.11
388	7	9.1	10	13	5.5	8.6	12.5	7.9	8.4	4.9	4.9	10.1	5.3	5	6.6	7.5	6.6	5.3	5.7	5.6
389	1.94	-19.92	-9.68	-10.95	-1.24	-9.38	-10.2	2.39	-10.27	2.15	2.28	-9.52	-1.48	-0.76	-3.68	-5.06	-4.88	-5.88	-6.11	1.99
390	0.07	2.88	3.22	3.64	0.71	2.18	3.08	2.23	2.41	-4.44	-4.19	2.84	-2.49	-4.92	-1.22	1.96	0.92	-4.75	-1.39	-2.69
391	-1.73	2.52	1.45	1.13	-0.97	0.53	0.39	-5.36	1.74	-1.68	-1.03	1.41	-0.27	1.3	0.88	-1.63	-2.09	3.65	2.32	-2.53
392	0.09	-3.44	0.84	2.36	4.13	-1.14	-0.07	0.3	1.11	-1.03	-0.98	-3.14	-0.41	0.45	2.23	0.57	-1.4	0.85	0.01	-1.29
393	8.5	0	8.2	8.5	11	6.3	8.8	7.1	10.1	16.8	15	7.9	13.3	11.2	8.2	7.4	8.8	9.9	8.8	12
394	6.8	0	6.2	7	8.3	8.5	4.9	6.4	9.2	10	12.2	7.5	8.4	8.3	6.9	8	7	5.7	6.8	9.4
395	18.08	0	17.47	17.36	18.17	17.93	18.16	18.24	18.49	18.62	18.6	17.96	18.11	17.3	18.16	17.57	17.54	17.19	17.99	18.3
396	18.56	0	18.24	17.94	17.84	18.51	17.97	18.57	18.64	19.21	19.01	18.36	18.49	17.95	18.77	18.06	17.71	16.87	18.23	18.98
397	-0.152	-0.089	-0.203	-0.355	0	-0.181	-0.411	-0.19	0	-0.086	-0.102	-0.062	-0.107	0.001	-0.181	-0.203	-0.17	0.275	0	-0.125
398	0.83	0.83	0.09	0.64	1.48	0	0.65	0.1	1.1	3.07	2.52	1.6	1.4	2.75	2.7	0.14	0.54	0.31	2.97	1.79
399	11.5	14.28	12.82	11.68	13.46	14.45	13.57	3.4	13.69	21.4	21.4	15.71	16.25	19.8	17.43	9.47	15.77	21.67	18.03	21.57
400	0	52	3.38	49.7	1.48	3.53	49.9	0	51.6	0.13	0.13	49.5	1.43	0.35	1.58	1.67	1.66	2.1	1.61	0.13
401	6	10.76	5.41	2.77	5.05	5.65	3.22	5.97	7.59	6.02	5.98	9.74	5.74	5.48	6.3	5.68	5.66	5.89	5.66	5.96
402	9.9	4.6	5.4	2.8	2.8	9	3.2	5.6	8.2	17.1	17.6	3.5	14.9	18.8	14.8	6.9	9.5	17.1	15	14.3
403	0.94	1.15	0.79	1.19	0.6	0.94	1.41	1.18	1.15	1.07	0.95	1.03	0.88	1.06	1.18	0.69	0.87	0.91	1.04	0.9
404	0.98	1.14	1.05	1.05	0.41	0.9	1.04	1.25	1.01	0.88	0.8	1.06	1.12	1.12	1.31	1.02	0.8	0.9	1.12	0.87
405	1.05	0.81	0.91	1.39	0.6	0.87	1.11	1.26	1.43	0.95	0.96	0.97	0.99	0.95	1.05	0.96	1.03	1.06	0.94	0.62
406	0.75	0.9	1.24	1.72	0.66	1.08	1.1	1.14	0.96	0.8	1.01	0.66	1.02	0.88	1.33	1.2	1.13	0.68	0.8	0.58

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
407	0.67	0.76	1.28	1.58	0.37	1.05	0.94	0.98	0.83	0.78	0.79	0.84	0.98	0.96	1.12	1.25	1.41	0.94	0.82	0.67
408	1.1	1.05	0.72	1.14	0.26	1.31	2.3	0.55	0.83	1.06	0.84	1.08	0.9	0.9	1.67	0.81	0.77	1.26	0.99	0.76
409	1.39	0.95	0.67	1.64	0.52	1.6	2.07	0.65	1.36	0.64	0.91	0.8	1.1	1	0.94	0.69	0.92	1.1	0.73	0.7
410	1.43	1.33	0.55	0.9	0.52	1.43	1.7	0.56	0.66	1.18	1.52	0.82	1.68	1.1	0.15	0.61	0.75	1.68	0.65	1.14
411	1.55	1.39	0.6	0.61	0.59	1.43	1.34	0.37	0.89	1.47	1.36	1.27	2.13	1.39	0.03	0.44	0.65	1.1	0.93	1.18
412	1.8	1.73	0.73	0.9	0.55	0.97	1.73	0.32	0.46	1.09	1.47	1.24	1.64	0.96	0.15	0.67	0.7	0.68	0.91	0.81
413	1.52	1.49	0.58	1.04	0.26	1.41	1.76	0.3	0.83	1.25	1.26	1.1	1.14	1.14	0.44	0.66	0.73	0.68	1.04	1.03
414	1.49	1.41	0.67	0.94	0.37	1.52	1.55	0.29	0.96	1.04	1.4	1.17	1.84	0.86	0.2	0.68	0.79	1.52	1.06	0.94
415	1.73	1.24	0.7	0.68	0.63	0.88	1.16	0.32	0.76	1.15	1.8	1.22	2.21	1.35	0.07	0.65	0.46	1.57	1.1	0.94
416	1.33	1.39	0.64	0.6	0.44	1.37	1.43	0.2	1.02	1.58	1.63	1.71	1.76	1.22	0.07	0.42	0.57	1	1.02	1.08
417	1.87	1.66	0.7	0.91	0.33	1.24	1.88	0.33	0.89	0.9	1.65	1.63	1.35	0.67	0.03	0.71	0.5	1	0.73	0.51
418	1.19	1.45	1.33	0.72	0.44	1.43	1.27	0.74	1.55	0.61	1.36	1.45	1.35	1.2	0.1	1.02	0.82	0.58	1.06	0.46
419	0.77	1.11	1.39	0.79	0.44	0.95	0.92	2.74	1.65	0.64	0.66	1.19	0.74	1.04	0.66	0.64	0.82	0.58	0.93	0.53
420	0.93	0.96	0.82	1.15	0.67	1.02	1.07	1.08	1.4	1.14	1.16	1.27	1.11	1.05	1.01	0.71	0.84	1.06	1.15	0.74
421	1.09	1.29	1.03	1.17	0.26	1.08	1.31	0.97	0.88	0.97	0.87	1.13	0.96	0.84	2.01	0.76	0.79	0.91	0.64	0.77
422	0.71	1.09	0.95	1.43	0.65	0.87	1.19	1.07	1.13	1.05	0.84	1.1	0.8	0.95	1.7	0.65	0.086	1.25	0.85	1.12
423	13.4	13.3	12	11.7	11.6	12.8	12.2	11.3	11.6	12	13	13	12.8	12.1	6.5	12.2	11.7	12.4	12.1	11.9
424	-0.77	-0.68	-0.07	-0.15	-0.23	-0.33	-0.27	0	-0.06	-0.23	-0.62	-0.65	-0.5	-0.41	3	-0.35	-0.11	-0.45	-0.17	-0.14
425	0.984	1.008	1.048	1.068	0.906	1.037	1.094	1.031	0.95	0.927	0.935	1.102	0.952	0.915	1.049	1.046	0.997	0.904	0.929	0.931
426	1.315	1.31	1.38	1.372	1.196	1.342	1.376	1.382	1.279	1.241	1.234	1.367	1.269	1.247	1.342	1.381	1.324	1.186	1.199	1.235
427	0.994	1.026	1.022	1.022	0.939	1.041	1.052	1.018	0.967	0.977	0.982	1.029	0.963	0.934	1.05	1.025	0.998	0.938	0.981	0.968
428	0.783	0.807	0.799	0.822	0.785	0.817	0.826	0.784	0.777	0.776	0.783	0.834	0.806	0.774	0.809	0.811	0.795	0.796	0.788	0.781
429	0.423	0.503	0.906	0.87	0.877	0.594	0.167	1.162	0.802	0.566	0.494	0.615	0.444	0.706	1.945	0.928	0.884	0.69	0.778	0.706
430	0.619	0.753	1.089	0.932	1.107	0.77	0.675	1.361	1.034	0.876	0.74	0.784	0.736	0.968	1.78	0.969	1.053	0.91	1.009	0.939
431	1.08	0.976	1.197	1.266	0.733	1.05	1.085	1.104	0.906	0.583	0.789	1.026	0.812	0.685	1.412	0.987	0.784	0.755	0.665	0.546
432	0.978	0.784	0.915	1.038	0.573	0.863	0.962	1.405	0.724	0.502	0.766	0.841	0.729	0.585	2.613	0.784	0.569	0.671	0.56	0.444
433	1.4	1.23	1.61	1.89	1.14	1.33	1.42	2.06	1.25	1.02	1.33	1.34	1.12	1.07	3.9	1.2	0.99	1.1	0.98	0.87
434	4.08	3.91	3.83	3.02	4.49	3.67	2.23	4.24	4.08	4.52	4.81	3.77	4.48	5.38	3.8	4.12	4.11	6.1	5.19	4.18
435	-0.35	-0.44	-0.38	-0.41	-0.47	-0.4	-0.41	0	-0.46	-0.56	-0.48	-0.41	-0.46	-0.55	-0.23	-0.39	-0.48	-0.48	-0.5	-0.53
436	0.5	1.7	1.7	1.6	0.6	1.6	1.6	1.3	1.6	0.6	0.4	1.6	0.5	0.4	1.7	0.7	0.4	0.7	0.6	0.5
437	0.96	0.77	0.39	0.42	0.42	0.8	0.53	0	0.57	0.84	0.92	0.73	0.86	0.59	-2.5	0.53	0.54	0.58	0.72	0.63
438	0.343	0.353	0.409	0.429	0.319	0.395	0.405	0.389	0.307	0.296	0.287	0.429	0.293	0.292	0.432	0.416	0.362	0.268	0.22	0.307
439	0.32	0.327	0.384	0.424	0.198	0.436	0.514	0.374	0.299	0.306	0.34	0.446	0.313	0.314	0.354	0.376	0.339	0.291	0.287	0.294
440	8.9	4.6	4.4	6.3	0.6	2.8	6.9	9.4	2.2	7	7.4	6.1	2.3	3.3	4.2	4	5.7	1.3	4.5	8.2
441	9.2	3.6	5.1	6	1	2.9	6	9.4	2.1	6	7.7	6.5	2.4	3.4	4.2	5.5	5.7	1.2	3.7	8.2
442	14.1	5.5	3.2	5.7	0.1	3.7	8.8	4.1	2	7.1	9.1	7.7	3.3	5	0.7	3.9	4.4	1.2	4.5	5.9
443	13.4	3.9	3.7	4.6	0.8	4.8	7.8	4.6	3.3	6.5	10.6	7.5	3	4.5	1.3	3.8	4.6	1	3.3	7.1

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
444	9.8	7.3	3.6	4.9	3	2.4	4.4	0	11.9	17.2	17	10.5	11.9	23	15	2.6	6.9	24.2	17.2	15.3
445	0.7	0.95	1.47	0.87	1.17	0.73	0.96	0.64	1.39	1.29	1.44	0.91	0.91	1.34	0.12	0.84	0.74	1.8	1.68	1.2
446	58	-184	-93	-97	116	-139	-131	-11	-73	107	95	-24	78	92	-79	-34	-7	59	-11	100
447	51	-144	-84	-78	137	-128	-115	-13	-55	106	103	-205	73	108	-79	-26	-3	69	11	108
448	41	-109	-74	-47	169	-104	-90	-18	-35	104	103	-148	77	128	-81	-31	10	102	36	116
449	32	-95	-73	-29	182	-95	-74	-22	-25	106	104	-124	82	132	-82	-34	20	118	44	113
450	24	-79	-76	0	194	-87	-57	-28	-31	102	103	-9	90	131	-85	-36	34	116	43	111
451	5	-57	-77	45	224	-67	-8	-47	-50	83	82	-38	83	117	-103	-41	79	130	27	117
452	-2	-41	-97	248	329	-37	117	-66	-70	28	36	115	62	120	-132	-52	174	179	-7	114
453	0.4	1.5	1.6	15	0.7	1.4	1.3	1.1	1.4	0.5	0.3	1.4	0.5	0.3	1.6	0.9	0.7	0.9	0.9	0.4
454	-0.04	-0.3	0.25	0.27	0.57	-0.02	-0.33	1.24	-0.11	-0.26	-0.38	-0.18	-0.09	-0.01	0	0.15	0.39	0.21	0.05	-0.06
455	-0.12	0.34	1.05	1.12	-0.63	1.67	0.91	0.76	1.34	-0.77	0.15	0.29	-0.71	-0.67	0	1.45	-0.7	-0.14	-0.49	-0.7
456	8.6	4.2	4.6	4.9	2.9	4	5.1	7.8	2.1	4.6	8.8	6.3	2.5	3.7	4.9	7.3	6	1.4	3.6	6.7
457	7.6	5	4.4	5.2	2.2	4.1	6.2	6.9	2.1	5.1	9.4	5.8	2.1	4	5.4	7.2	6.1	1.4	3.2	6.7
458	8.1	4.6	3.7	3.8	2	3.1	4.6	7	2	6.7	11	4.4	2.8	5.6	4.7	7.3	5.6	1.8	3.3	7.7
459	7.9	4.9	4	5.5	1.9	4.4	7.1	7.1	2.1	5.2	8.6	6.7	2.4	3.9	5.3	6.6	5.3	1.2	3.1	6.8
460	8.3	8.7	3.7	4.7	1.6	4.7	6.5	6.3	2.1	3.7	7.4	7.9	2.3	2.7	6.9	8.8	5.1	0.7	2.4	5.3
461	4.47	8.48	3.89	7.05	0.29	2.87	16.56	8.29	1.74	3.3	5.06	12.98	1.71	2.32	5.41	4.27	3.83	0.67	2.75	4.05
462	6.77	6.87	5.5	8.57	0.31	5.24	12.93	7.95	2.8	2.72	4.43	10.2	1.87	1.92	4.79	5.41	5.36	0.54	2.26	3.57
463	7.43	4.51	9.12	8.71	0.42	5.42	5.86	9.4	1.49	1.76	2.74	9.67	0.6	1.18	5.6	9.6	8.95	1.18	3.26	3.1
464	5.22	7.3	6.06	7.91	1.01	6	10.66	5.81	2.27	2.36	4.52	12.68	1.85	1.68	5.7	6.99	5.16	0.56	2.16	4.1
465	9.88	3.71	2.35	3.5	1.12	1.66	4.02	6.88	1.88	10.08	13.21	3.39	2.44	5.27	3.8	4.1	4.98	1.11	4.07	12.53
466	10.98	3.26	2.85	3.37	1.47	2.3	3.51	7.48	2.2	9.74	12.79	2.54	3.1	4.97	3.42	4.93	5.55	1.28	3.55	10.69
467	9.95	3.05	4.84	4.46	1.3	2.64	2.58	8.87	1.99	7.73	9.66	2	2.45	5.41	3.2	6.03	5.62	2.6	6.15	9.46
468	8.26	2.8	2.54	2.8	2.67	2.86	2.67	5.62	1.98	8.95	16.46	1.89	2.67	7.32	3.3	6	5	2.01	3.96	10.24
469	7.39	5.91	3.06	5.14	0.74	2.22	9.8	7.53	1.82	6.96	9.45	7.81	2.1	3.91	4.54	4.18	4.45	0.9	3.46	8.62
470	9.07	4.9	4.05	5.73	0.95	3.63	7.77	7.69	2.47	6.56	9	6.01	2.54	3.59	4.04	5.15	5.46	0.95	2.96	7.47
471	8.82	3.71	6.77	6.38	0.9	3.89	4.05	9.11	1.77	5.05	6.54	5.45	1.62	3.51	4.28	7.64	7.12	1.96	4.85	6.6
472	6.65	5.17	4.4	5.5	1.79	4.52	6.89	5.72	2.13	5.47	10.15	7.59	2.24	4.34	4.56	6.52	5.08	1.24	3.01	7
473	0	2.45	0	0	0	1.25	1.27	0	1.45	0	0	3.67	0	0	0	0	0	6.93	5.06	0
474	89.3	190.3	122.4	114.4	102.5	146.9	138.8	63.8	157.5	163	163.1	165.1	165.8	190.8	121.6	94.2	119.6	226.4	194.6	138.2
475	90	194	124.7	117.3	103.3	149.4	142.2	64.9	160	163.9	164	167.3	167	191.9	122.9	95.4	121.5	228.2	197	139
476	0.0373	0.0959	0.0036	0.1263	0.0829	0.0761	0.0058	0.005	0.0242	0	0	0.0371	0.0823	0.0946	0.0198	0.0829	0.0941	0.0548	0.0516	0.0057
477	0.85	0.2	-0.48	-1.1	2.1	-0.42	-0.79	0	0.22	3.14	1.99	-1.19	1.42	1.69	-1.14	-0.52	-0.08	1.76	1.37	2.53
478	0.06	-0.85	0.25	-0.2	0.49	0.31	-0.1	0.21	-2.24	3.48	3.5	-1.62	0.21	4.8	0.71	-0.62	0.65	2.29	1.89	1.59
479	2.62	1.26	-1.27	-2.84	0.73	-1.69	-0.45	-1.15	-0.74	4.38	6.57	-2.78	-3.12	9.14	-0.12	-1.39	1.81	5.91	1.39	2.3
480	-1.64	-3.28	0.83	0.7	9.3	-0.04	1.18	-1.85	7.17	3.02	0.83	-2.36	4.26	-1.36	3.12	1.59	2.31	2.61	2.37	0.52

Continued on next page

Table B.1 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
481	-2.34	1.6	2.81	-0.48	5.03	0.16	1.3	-1.06	-3	7.26	1.09	1.56	0.62	2.57	-0.15	1.93	0.19	3.59	-2.58	2.06
482	0.78	1.58	1.2	1.35	0.55	1.19	1.45	0.68	0.99	0.47	0.56	1.1	0.66	0.47	0.69	1	1.05	0.7	1	0.51
483	25	-7	-7	2	32	0	14	-2	-26	91	100	-26	68	100	25	-2	7	109	56	62
484	1.1	-5.1	-3.5	-3.6	2.5	-3.68	-3.2	-0.64	-3.2	4.5	3.8	-4.11	1.9	2.8	-1.9	-0.5	-0.7	-0.46	-1.3	4.2
485	0.137	0.036	-0.035	-0.123	0.275	0.033	-0.048	-0.047	0.055	0.417	0.425	-0.010	0.178	0.408	0.002	-0.043	0.059	0.236	0.317	0.408
486	0.073	0.039	-0.039	-0.055	0.356	0.013	-0.030	-0.059	0.087	0.381	0.382	-0.005	0.161	0.420	-0.049	-0.028	0.024	0.411	0.311	0.295
487	0.151	-0.010	0.038	0.005	0.322	0.025	-0.064	0.025	0.134	0.424	0.393	-0.016	0.216	0.346	0.084	0.004	0.146	0.266	0.230	0.400
488	-0.058	0	0.027	0.016	0.447	-0.073	-0.128	0.331	0.195	0.06	0.138	-0.112	0.275	0.24	-0.478	-0.177	-0.163	0.564	0.322	-0.052
489	-0.17	0.37	0.18	0.37	-0.06	0.26	0.15	0.01	-0.02	-0.28	-0.28	0.32	-0.26	-0.41	0.13	0.05	0.02	-0.15	-0.09	-0.17
490	-0.15	0.32	0.22	0.41	-0.15	0.03	0.3	0.08	0.06	-0.29	-0.36	0.24	-0.19	-0.22	0.15	0.16	-0.08	-0.28	-0.03	-0.24
491	0.964	1.143	0.944	0.916	0.778	1.047	1.051	0.835	1.014	0.922	1.085	0.944	1.032	1.119	1.299	0.947	1.017	0.895	1	0.955
492	0.974	1.129	0.988	0.892	0.972	1.092	1.054	0.845	0.949	0.928	1.11	0.946	0.923	1.122	1.362	0.932	1.023	0.879	0.902	0.923
493	0.938	1.137	0.902	0.857	0.6856	0.916	1.139	0.892	1.109	0.986	1	0.952	1.077	1.11	1.266	0.956	1.018	0.971	1.157	0.959
494	1.042	1.069	0.828	0.97	0.5	1.111	0.992	0.743	1.034	0.852	1.193	0.979	0.998	0.981	1.332	0.984	0.992	0.96	1.12	1.001
495	1.065	1.131	0.762	0.836	1.015	0.861	0.736	1.022	0.973	1.189	1.192	0.478	1.369	1.368	1.241	1.097	0.822	1.017	0.836	1.14
496	0.99	1.132	0.873	0.915	0.644	0.999	1.053	0.785	1.054	0.95	1.106	1.003	1.093	1.121	1.314	0.911	0.988	0.939	1.09	0.957
497	0.892	1.154	1.144	0.925	1.035	1.2	1.115	0.917	0.992	0.817	0.994	0.944	0.782	1.058	1.309	0.986	1.11	0.841	0.866	0.9
498	1.092	1.239	0.927	0.919	0.662	1.124	1.199	0.698	1.012	0.912	1.276	1.008	1.171	1.09	0.8	0.886	0.832	0.981	1.075	0.908
499	0.843	1.038	0.956	0.906	0.896	0.968	0.9	0.978	1.05	0.946	0.885	0.893	0.878	1.151	1.816	1.003	1.189	0.852	0.945	0.999
500	2.18	2.71	1.85	1.75	3.89	2.16	1.89	1.17	2.51	4.5	4.71	2.12	3.63	5.88	2.09	1.66	2.18	6.46	5.01	3.77
501	1.79	3.2	2.83	2.33	2.22	2.37	2.52	0.7	3.06	4.59	4.72	2.5	3.91	4.84	2.45	1.82	2.45	5.64	4.46	3.67
502	13.4	8.5	7.6	8.2	22.6	8.5	7.3	7	11.3	20.3	20.8	6.1	15.7	23.9	9.9	8.2	10.3	24.5	19.5	19.5
503	0.017	-0.076	-0.079	-0.128	0.572	-0.105	-0.180	-0.044	0.164	0.276	0.252	-0.213	0.020	0.356	-0.419	-0.163	-0.070	0.384	0.25	0.178
504	90.1	192.8	127.5	117.1	113.2	149.4	140.8	63.8	159.3	164.9	164.6	170	167.7	193.5	123.1	94.2	120	197.1	231.7	139.1
505	91.5	196.1	138.3	135.2	114.4	156.4	154.6	67.5	163.2	162.6	163.4	162.5	165.9	198.8	123.4	102	126	209.8	237.2	138.4
506	1.076	1.361	1.056	1.29	0.753	0.729	1.118	1.346	0.985	0.926	1.054	1.105	0.974	0.869	0.82	1.342	0.871	0.666	0.531	1.131
507	0.616	0	0.236	0.028	0.68	0.251	0.043	0.501	0.165	0.943	0.943	0.283	0.738	1	0.711	0.359	0.45	0.878	0.88	0.825

TABLE B.2: Real-values of amino acid indices with unknown descriptions.

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
508	0.24	3.52	3.05	3.98	0.84	1.75	3.11	2.05	2.47	-3.89	-4.28	2.29	-2.85	-4.22	-1.66	2.39	0.75	-4.36	-2.54	-2.59
509	-2.32	2.5	1.62	0.93	-1.67	0.5	0.26	-4.06	1.95	-1.73	-1.3	0.89	-0.22	1.94	0.27	-1.07	-2.18	3.94	2.44	-2.64
510	0.6	-3.5	1.04	1.93	3.71	-1.44	-0.11	0.36	0.26	-1.71	-1.49	-2.49	0.47	1.06	1.84	1.15	-1.12	0.59	0.43	-1.54
511	-0.14	1.99	-1.15	-2.46	0.18	-1.34	-3.04	-0.82	3.9	-0.84	-0.72	1.49	1.94	0.54	0.7	-1.39	-1.46	3.44	0.04	-0.85
512	1.3	-0.17	1.61	0.75	-2.65	0.66	-0.25	-0.38	0.09	0.26	0.84	0.31	-0.98	-0.62	2	0.67	-0.4	-1.59	-1.47	-0.02
513	60	1	39	66	35	53	74	44	9	80	82	1	77	82	52	44	58	86	81	74
514	24	9	14	16	23	16	24	19	6	50	52	7	45	50	20	19	25	54	50	42
515	29	6	25	9	27	37	15	22	16	49	50	2	50	52	50	15	27	54	47	43
516	9	1	5	2	40	7	4	6	1	28	33	1	29	35	11	6	11	40	32	22
517	52	19	21	38	46	32	42	32	16	72	73	19	63	66	57	36	46	49	68	68
518	23	7	12	20	63	20	19	22	20	53	53	7	49	52	31	23	39	51	51	43
519	37	1	12	19	81	23	28	17	2	83	83	1	79	87	68	13	30	88	80	72
520	4.13	4.06	4.33	4.34	4.3	4.1	4.12	3.88	4.29	4.02	4.05	4.05	4.2	4.31	4.37	4.18	3.97	4.36	4.3	3.94
521	3.77	3.76	4	3.89	3.97	3.76	3.74	3.54	4	3.66	3.71	3.74	3.84	3.98	4.11	3.87	3.57	4.05	3.9	3.6
522	3.3	3.2	3.56	3.51	3	3.24	3.21	3.15	3.47	3.08	3.23	3.27	3.3	3.48	3.51	3.31	3.08	3.56	3.4	3.03
523	0.008	0.171	0.255	0.303	-0.132	0.149	0.221	0.218	0.023	-0.353	-0.267	0.243	-0.239	-0.329	0.173	0.199	0.068	-0.296	-0.141	-0.274
524	0.134	-0.361	0.038	-0.057	0.174	-0.184	-0.28	0.562	-0.177	0.071	0.018	-0.339	-0.141	-0.023	0.286	0.238	0.147	-0.186	-0.057	0.136
525	-0.475	0.107	0.117	-0.014	0.07	-0.03	-0.315	-0.024	0.041	-0.088	-0.265	-0.044	-0.155	0.072	0.407	-0.015	-0.015	0.389	0.425	-0.187
526	-0.039	-0.258	0.118	0.225	0.565	0.035	0.157	0.018	0.28	-0.195	-0.274	-0.325	0.321	-0.002	-0.215	-0.068	-0.132	0.083	-0.096	-0.196
527	0.181	-0.364	-0.055	0.156	-0.374	-0.112	0.303	0.106	-0.021	-0.107	0.206	-0.027	0.077	0.208	0.384	-0.196	-0.274	0.297	-0.091	-0.299
528	0.354	7.573	11.294	13.42	-5.846	6.599	9.788	9.655	1.019	-15.634	-11.825	10.762	-10.585	-14.571	7.662	8.813	3.012	-13.11	-6.245	-12.135
529	3.762	-10.135	1.067	-1.6	4.885	-5.166	-7.861	15.778	-4.969	1.993	0.505	-9.517	-3.959	-0.646	8.029	6.682	4.127	-5.222	-1.6	3.818
530	-11.036	2.486	2.718	-0.325	1.626	-0.697	-7.318	-0.558	0.953	-2.045	-6.157	-1.022	-3.601	1.673	9.456	-0.348	-0.348	9.038	9.874	-4.345
531	-0.649	-4.291	1.963	3.742	9.397	0.582	2.611	0.299	4.657	-3.243	-4.557	-5.405	5.339	-0.033	-3.576	-1.131	-2.195	1.38	-1.597	-3.26
532	2.828	-5.687	-0.859	2.437	-5.843	-1.75	4.734	1.656	-0.328	-1.672	3.219	-0.422	1.203	3.25	6	-3.062	-4.281	4.64	-1.422	-4.672
533	0.5	0.74	0.78	1.33	0.53	0.82	1.26	0.75	0.69	0.47	0.45	0.55	0.48	0.47	0.65	0.7	0.68	0.58	0.79	0.45
534	0.43	1.21	0.83	0.71	0.39	0.72	0.69	0.62	0.89	0.38	0.33	1.27	0.41	0.4	0.58	0.81	0.7	0.55	0.65	0.37
535	0.46	0.99	0.8	0.99	0.46	0.77	0.95	0.68	0.79	0.42	0.38	0.94	0.44	0.43	0.61	0.76	0.69	0.57	0.72	0.41
536	0.79	1.09	1.12	1.18	0.77	1.04	1.2	1.03	1.1	0.68	0.67	1.15	0.74	0.71	1	1.04	0.94	0.8	0.83	0.69
537	1.79	1.04	1.1	0.95	1.53	1.24	1.06	1.39	1.21	1.96	2.02	1.09	1.84	1.88	1.43	1.21	1.42	1.67	1.52	1.94
538	1.83	1.56	1.57	1.47	1.71	1.66	1.61	1.84	1.73	1.82	1.93	1.58	2	1.88	1.97	1.57	1.58	1.74	1.58	1.78
539	0.81	0.74	1.08	1.87	0.68	1.12	1.95	1.29	0.94	0.94	0.95	0.64	0.94	0.93	0.98	1.13	1.1	1.03	1.12	0.97
540	0.75	1.62	1.1	0.73	0.63	0.96	0.65	0.96	1.08	0.77	0.7	1.65	0.72	0.77	0.9	1.1	1.06	0.83	0.94	0.76
541	0.78	1.21	1.09	1.26	0.66	1.03	1.25	1.11	1.01	0.85	0.82	1.18	0.82	0.84	0.94	1.11	1.08	0.92	1.02	0.86
542	0.99	1	1.13	1.15	0.94	1.07	1.09	1.16	1.03	0.87	0.86	1	0.89	0.93	1.06	1.07	1.05	0.94	0.9	0.89
543	1.25	1.02	0.9	0.81	0.99	1.05	0.93	1.02	1.05	1.33	1.41	1.03	1.33	1.32	1.12	0.91	1.01	1.25	1.19	1.33

Continued on next page

Table B.2 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
544	1.61	1.44	1.28	1.24	1.34	1.44	1.38	1.59	1.58	1.45	1.48	1.48	1.68	1.53	1.66	1.26	1.18	1.48	1.33	1.36
545	1.05	0.78	1.35	2.02	0.91	1.21	2.08	1.57	1.07	1.13	1	0.71	0.97	1.04	1.11	1.13	1.19	1	1.1	1.15
546	1.33	1.95	1.4	0.93	0.91	1.21	0.87	1.24	1.5	1.15	0.92	1.96	0.88	0.85	1.17	1.62	1.48	1.11	1.11	1.05
547	1.2	1.39	1.38	1.43	0.91	1.21	1.42	1.39	1.3	1.14	0.96	1.36	0.93	0.94	1.14	1.39	1.35	1.06	1.11	1.1
548	1.21	1.02	1.13	1.21	1.09	1.13	1.15	1.3	1.09	1.13	1.08	1.11	1.04	1.1	1.2	1.19	1.16	1.05	1.04	1.18
549	0.66	0.81	0.64	0.6	0.64	0.83	0.69	0.74	0.7	0.86	1.05	0.73	0.99	0.95	0.83	0.56	0.61	0.94	0.92	0.81
550	0.84	1.14	0.9	0.9	0.81	1.08	0.99	1.15	1.27	0.99	1.21	1.08	1.17	1.33	1.24	0.77	0.72	1.26	1.24	0.91
551	0.63	0.75	1.09	1.79	0.57	1.08	1.87	1.03	0.85	0.59	0.58	0.67	0.62	0.61	0.89	0.93	0.92	0.75	0.94	0.6
552	0.61	1.62	1.12	0.81	0.44	0.99	0.76	0.81	1.07	0.49	0.43	1.78	0.51	0.5	0.85	1.11	0.99	0.67	0.8	0.49
553	0.62	1.21	1.11	1.26	0.5	1.03	1.27	0.91	0.97	0.54	0.5	1.25	0.56	0.55	0.87	1.03	0.95	0.71	0.86	0.54
554	0.88	1.03	1.13	1.18	0.81	1.08	1.14	1.11	1.08	0.74	0.73	1.08	0.8	0.78	1.08	1.09	1.02	0.86	0.87	0.76
555	1.53	0.96	0.86	0.76	1.43	1.01	0.84	1.18	1.05	1.78	1.84	0.88	1.67	1.71	1.16	0.95	1.11	1.5	1.34	1.75
556	1.65	1.38	1.23	1.16	1.62	1.37	1.25	1.64	1.59	1.71	1.8	1.28	1.87	1.77	1.67	1.26	1.26	1.63	1.46	1.63
557	0.71	1.09	0.95	1.43	0.65	0.87	1.19	1.07	1.13	1.05	0.84	1.1	0.8	0.95	1.7	0.65	0.86	1.25	0.85	1.12
558	1.07	1.32	0.94	0.75	0.62	0.9	0.7	1.04	0.99	1	1.01	1	1.42	1.21	0.93	0.99	1	1.66	1.15	0.93
559	0.99	1.29	0.99	0.66	0.66	0.9	0.64	1.03	1	1.06	1.07	0.97	1.08	1.35	0.9	0.99	1.1	1.85	1.18	0.93
560	1.04	1.3	1	0.63	0.7	0.88	0.61	0.98	0.99	1.05	1.1	0.94	1.09	1.24	0.91	0.97	1.08	1.77	1.47	0.96
561	0.92	1.7	1.12	0.8	0.79	0.89	0.64	1.05	1.1	0.83	0.89	1.11	1.13	0.99	0.91	1.01	1.15	1.83	1.34	0.79
562	0.79	2.38	1.44	1.15	0.74	1.21	0.93	0.87	1.43	0.3	0.53	1.6	0.73	0.4	0.94	1.2	1.23	1.04	1.14	0.42
563	1.31	0.15	0.28	0.1	0.64	0.18	0.1	0.94	0.3	2.72	2.02	0.09	1.64	2.59	0.51	0.41	0.65	2.87	1.37	1.99
564	1.43	0.13	0.37	0.14	0.79	0.26	0.15	1.11	0.38	2.19	1.9	0.12	1.59	2.33	0.51	0.7	0.8	2.51	1.42	1.76
565	1.36	0.17	0.43	0.22	0.81	0.36	0.19	1.14	0.63	2.11	1.77	0.17	1.81	2.17	0.6	0.72	0.86	2.1	1.38	1.66
566	1.49	0.16	0.35	0.17	0.87	0.27	0.13	1.22	0.56	2.11	1.67	0.12	1.54	2.21	0.6	0.78	0.94	2.09	1.46	1.75
567	1.52	0.16	0.42	0.15	1.05	0.32	0.14	1.14	0.44	2.19	1.91	0.12	1.49	2.06	0.56	0.72	1.01	1.45	1.03	1.78
568	1.383	0.124	0.389	0.153	1.202	0.273	0.131	1.158	0.395	2.083	1.845	0.108	1.502	2.235	0.597	0.806	0.879	1.79	1.075	1.756
569	0.324	-2.085	-0.944	-1.877	0.184	-1.3	-2.033	0.147	-0.93	0.734	0.612	-2.23	0.407	0.804	-0.516	-0.216	-0.129	0.582	0.073	0.563
570	97	150	103	95	78	119	110	64	144	170	171	109	182	193	95	87	107	226	192	146
571	-0.04	-0.02	0	0.01	0	-0.02	-0.03	0.03	0	0.01	-0.01	-0.01	0.01	0.01	0.08	0.01	0.02	0	0.01	0.01
572	-0.06	-0.03	0.12	0.22	0	-0.03	-0.03	0.06	0	-0.03	-0.06	-0.03	-0.01	-0.02	0.06	0.25	0.15	-0.01	-0.01	-0.03
573	-0.01	-0.01	-0.02	-0.01	0	-0.01	0.02	-0.01	-0.01	-0.01	-0.02	0	-0.01	0	0.26	-0.02	-0.01	0	0	0.01
574	0.01	-0.01	-0.01	0.08	0	0.01	0.14	-0.01	0	-0.03	-0.06	0.01	-0.01	-0.01	0.05	0.02	-0.01	0.01	-0.01	-0.02
575	-0.02	-0.02	-0.02	0.05	0	0.04	0.09	-0.02	0.01	-0.02	-0.03	-0.02	-0.01	0.01	0	0	0.01	0	-0.01	0
576	0.02	0.02	-0.02	-0.03	0	0	-0.04	-0.03	-0.01	0.07	0.06	-0.02	0.01	0.04	0	-0.03	-0.01	0.01	0.02	0.1
577	0.04	0.01	-0.02	-0.03	0.01	0	-0.01	-0.03	-0.01	0.03	0.07	-0.01	0.01	0.02	-0.01	-0.02	-0.01	0.02	0.02	0.01
578	0.05	0.01	0	-0.02	0	0	0	-0.03	0	0	0.07	0.02	0.03	0	-0.01	-0.01	-0.02	0	0	-0.01
579	0.01	0.03	-0.01	-0.02	0	0.01	0.04	-0.03	0	0.01	0.02	0.07	0	-0.01	-0.01	-0.01	0	0	-0.01	0
580	0.03	0.01	0.01	-0.02	0	0	0.01	-0.03	0.01	-0.01	0.03	0.02	0	0.01	0	0.01	0.02	0	0.01	-0.02

Continued on next page

Table B.2 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
581	-0.01	-0.01	0.05	-0.01	0	0	-0.03	0.31	0.02	-0.03	-0.02	0	-0.01	0	0	0.01	-0.01	-0.01	0	-0.03
582	-0.05	0	0.02	-0.01	0	-0.01	-0.02	0.31	0	-0.02	-0.05	0.01	-0.01	-0.01	0.18	-0.01	-0.01	-0.01	-0.01	-0.02
583	-1.29	-0.13	-5.23	-6.55	-1.01	-5.3	-6.34	-1.36	-4.17	-1.03	-1.06	-0.13	-0.65	-0.8	-0.66	-4.36	-4.6	-0.74	-4.17	-0.84
584	-0.36	-5.46	-6.3	-6.77	-1.18	-5.98	-6.32	-1.13	-5.49	-0.86	-0.76	-5.5	-0.89	-0.99	-0.85	-5.49	-5.46	-5.04	-5.43	-0.76
585	-0.79	-5.21	-3.73	-1.52	-0.62	-3.4	-0.74	-0.2	-4.16	-0.58	-0.66	-5.29	-0.59	-1.09	-0.66	-4.27	-4.58	-4.61	-4.42	-0.66
586	-0.99	-7.98	-4.69	-1.49	-1.99	-4.95	-1.9	-2.36	-8.7	-1.64	-1.63	-6.34	-1.52	-2.34	-2.38	-2.75	-3.64	-6	-6.1	-1.2
587	-1.18	-2.99	-9.35	-9.39	-1.93	-9.3	-7.16	-2.85	-6.41	-2.04	-2.87	-2.38	-1.89	-1.42	-1.9	-2.64	-4.24	-4.67	-3.21	-2.09
588	-1.41	-0.33	-7.24	-13.84	-1.55	-10.02	-13.39	-0.46	-9.03	-0.76	-0.74	-0.34	-0.77	-1.2	-0.9	-4.14	-4.87	-3.85	-1.49	-1.3
589	445	606	492	483	474	532	529	413	544	529	540	559	545	590	483	454	472	649	600	490
590	242	278	295	314	233	307	340	250	256	227	224	302	219	218	236	300	287	273	250	226
591	245	322	263	261	250	290	289	225	283	273	291	299	277	298	278	254	256	313	303	253
592	435	485	494	532	417	529	582	449	445	401	398	539	391	388	421	515	491	416	463	404
593	92	182	145	169	120	158	168	91	153	128	127	147	129	161	98	97	110	214	169	114
594	146	273	230	269	182	251	273	135	230	191	190	228	183	237	221	162	179	261	252	165
595	-0.96	0.8	0.82	1	-0.55	0.78	0.94	-0.88	0.67	-0.94	-0.9	0.6	-0.82	-0.85	-0.81	0.41	0.4	0.06	0.31	-1
596	-0.76	0.63	-0.57	-0.89	-0.47	-0.3	-0.54	-1	-0.11	-0.05	0.03	0.1	0.03	0.48	-0.4	-0.82	-0.64	1	0.42	-0.43
597	0.31	0.99	0.02	-1	0.19	-0.38	-0.99	0.49	0.37	-0.18	-0.24	1	-0.08	-0.58	-0.07	0.57	0.37	-0.47	-0.2	-0.14
598	0.669	1.04	2.35	2.06	0.945	1.07	0.787	0.621	1.55	0.511	0.885	0.977	0.965	1.2	0.474	1.01	1.32	1.02	1.24	0.421
599	1.11	1.03	1.09	1.16	0.878	1.15	1.25	0.482	1.05	0.561	1.07	1.13	1.06	0.772	1.38	1.43	1.09	1.04	0.781	0.572
600	1.44	1.2	0.655	0.881	0.704	1.26	1.37	0.432	0.832	1.09	1.29	1.17	1.23	0.952	0.621	0.745	0.761	1.08	0.915	0.955
601	0.931	0.88	1.07	0.822	1.79	0.64	0.517	1.75	1.28	0.487	0.299	0.62	0.931	1.65	0.013	2.05	1.84	1.44	1.75	0.74
602	1.08	1.24	0.65	0.469	1.28	1.06	0.852	0.388	1.34	1.21	0.673	0.984	1.31	1.42	0.007	1.44	1.42	1.05	1.53	1.57
603	0.828	0.968	2.68	2.07	1.89	1.12	0.828	0.284	2.06	0.747	0.619	0.949	1.23	1.14	0.006	1.13	0.751	1.16	1.12	0.789
604	0.518	0.85	1.01	0.926	1.52	0.876	0.598	1.52	0.862	0.747	0.781	0.898	0.694	1.06	0.278	1.47	2.51	0.941	1.23	0.895
605	0.538	0.882	0.667	0.537	1.23	0.84	0.727	0.206	0.926	2.19	1.29	0.856	0.982	1.44	0.071	0.731	1.32	1.21	1.41	2.28
606	0.529	0.921	2	2.01	1.46	0.847	0.697	0.184	1.64	1.59	1.03	0.857	0.96	1.5	0.0997	0.635	0.746	1.09	1.24	1.37
607	0.762	0.578	0.899	1.25	1.13	0.609	0.484	1.73	0.629	0.237	0.521	0.61	0.524	0.465	4.38	2	1.48	0.542	0.521	0.275
608	0.996	0.846	0.791	1.05	1.21	0.783	0.866	0.297	0.849	0.873	0.945	0.949	0.842	0.829	3.82	1.02	0.792	1.02	0.848	0.887
609	0.744	0.812	2.13	2.51	1.38	0.778	0.881	0.395	1.05	0.795	0.879	0.904	0.815	0.841	2.4	0.725	0.442	0.828	0.669	0.73
610	0.307	0.582	2.53	1.17	0.5	0.655	0.473	7.07	1.04	0.049	0.198	0.785	0.33	0.465	0.006	0.475	0.152	0.277	0.402	0.065
611	0.288	0.256	0.266	0.321	0.226	0.176	0.204	10.9	0.312	0.0825	0.124	0.269	0.234	0.143	0.0205	0.436	0.192	0.15	0.147	0.0935
612	0.746	0.996	1.48	1.19	1.12	0.9	0.977	1.91	1.32	0.659	0.712	1.07	0.668	0.841	0.697	1.11	1.11	0.639	0.976	0.689
613	-0.99	0.28	0.77	0.74	0.34	0.12	0.59	-0.79	0.08	-0.77	-0.92	-0.63	-0.8	0.87	-0.99	0.99	0.42	-0.13	0.59	-0.99
614	-0.61	-0.99	-0.24	-0.72	0.88	-0.99	-0.55	-0.99	-0.71	0.67	0.31	0.25	0.44	0.65	-0.99	0.4	0.21	0.77	0.33	0.27
615	0	-0.22	0.59	-0.35	0.35	-0.99	-0.99	0.1	0.68	-0.37	-0.99	0.5	-0.71	-0.53	-0.99	0.37	0.97	-0.9	-0.99	-0.52
616	0.15	-1.47	-0.99	-1.15	0.18	-0.96	-1.18	-0.2	-0.43	1.27	1.36	-1.17	1.01	1.52	0.22	-0.67	-0.34	1.5	0.61	0.76
617	-1.11	1.45	0	0.67	-1.67	0.12	0.4	-1.53	-0.25	-0.14	0.07	0.7	-0.53	0.61	-0.17	-0.86	-0.51	2.06	1.6	-0.92

Continued on next page

Table B.2 – *Continued from previous page*

ID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
618	-1.35	1.24	-0.37	-0.41	-0.46	0.18	0.1	-2.63	0.37	0.3	0.26	0.7	0.43	0.96	-0.5	-1.07	-0.55	1.79	1.17	-0.17
619	-0.92	1.27	0.69	-0.01	-0.21	0.16	0.36	2.28	0.19	-1.8	-0.8	0.8	0	-0.16	0.05	-0.41	-1.06	0.75	0.73	-1.91
620	0.02	1.55	-0.55	-2.68	0	0.09	-2.16	-0.53	0.51	0.3	0.22	1.64	0.23	0.25	-0.01	-0.32	-0.06	0.75	0.53	0.22
621	-0.91	1.47	0.85	1.31	1.2	0.42	-0.17	-1.18	1.28	-1.61	-1.37	0.67	0.1	0.28	-1.34	0.27	-0.01	-0.13	0.25	-1.4
622	0.36	1.3	0.73	0.03	-1.61	-0.2	0.91	2.01	0.93	-0.16	0.08	1.63	-0.86	-1.33	-0.19	-0.64	-0.79	-1.01	-0.96	-0.24
623	-0.48	0.83	-0.8	0.56	-0.19	-0.41	0.02	-1.34	0.65	-0.13	-0.62	0.13	-0.68	-0.2	3.56	0.11	0.39	-0.85	-0.52	-0.03
624	0.34	-0.08	-0.16	0.04	-0.51	-0.16	0.03	0.19	-0.21	-0.45	-0.45	-0.2	-0.47	-0.47	0.17	0.16	0.18	-0.51	-0.34	-0.06
625	-0.08	0.65	0.01	-0.06	-0.35	0.15	-0.02	0.2	0.11	-0.82	-0.74	0.04	-0.83	-0.79	0.14	0.05	0	-0.26	-0.31	-0.54
626	-0.16	0.01	0.38	0.2	-0.46	0.2	0.19	-0.11	0.21	-0.83	-0.9	0.38	-0.97	-0.71	-0.12	-0.02	0.11	-0.79	0.12	-0.69
627	0.04	-0.06	0.2	0.36	-0.41	0.16	0.32	0.15	0.18	-0.78	-0.75	0.16	-0.9	-0.62	-0.15	-0.14	-0.1	-0.64	0.09	-0.39
628	-0.51	-0.35	-0.46	-0.41	1.02	-0.74	-0.64	-0.18	-0.29	-0.13	-0.02	-0.81	-0.29	0.39	-0.65	-0.2	-0.62	0.82	0.4	-0.19
629	-0.16	0.15	0.2	0.16	-0.74	0.48	0.25	-0.14	0.37	-0.95	-0.74	0.3	-0.97	-0.85	0.15	-0.2	-0.07	-0.83	-0.02	-0.68
630	0.03	-0.02	0.19	0.32	-0.64	0.25	0.36	0.13	0.13	-0.86	-0.8	0.25	-0.87	-0.81	-0.13	-0.19	-0.09	-0.66	-0.11	-0.42
631	0.19	0.2	-0.11	0.15	-0.18	-0.14	0.13	0.43	-0.19	-0.58	-0.56	-0.16	-0.61	-0.52	-0.14	0	-0.09	-0.15	-0.35	-0.15
632	-0.21	0.11	0.21	0.18	-0.29	0.37	0.13	-0.19	0.54	-0.69	-0.5	0.14	-0.86	-0.43	0.14	-0.15	-0.13	-0.7	0.35	-0.59
633	-0.45	-0.82	-0.83	-0.78	-0.13	-0.95	-0.86	-0.58	-0.69	0.75	0.48	-1.01	0.67	0.45	-0.78	-0.68	-0.59	-0.23	-0.35	0.41
634	-0.45	-0.74	-0.9	-0.75	-0.02	-0.74	-0.8	-0.56	-0.5	0.48	0.61	-1.06	0.5	0.48	-0.67	-0.7	-0.77	0.13	-0.27	0.41
635	-0.2	0.04	0.38	0.16	-0.81	0.3	0.25	-0.16	0.14	-1.01	-1.06	0.49	-1.03	-1.03	-0.14	-0.14	0.09	-0.92	-0.12	-0.79
636	-0.47	-0.83	-0.97	-0.9	-0.29	-0.97	-0.87	-0.61	-0.86	0.67	0.5	-1.03	0.97	0.35	-0.82	-0.75	-0.61	0.04	-0.56	0.41
637	-0.47	-0.79	-0.71	-0.62	0.39	-0.85	-0.81	-0.52	-0.43	0.45	0.48	-1.03	0.35	0.61	-0.76	-0.53	-0.75	0.25	0.14	0.36
638	0.17	0.14	-0.12	-0.15	-0.65	0.15	-0.13	-0.14	0.14	-0.78	-0.67	-0.14	-0.82	-0.76	0.56	0.24	0.25	-0.71	-0.2	-0.47
639	0.16	0.05	-0.02	-0.14	-0.2	-0.2	-0.19	0	-0.15	-0.68	-0.7	-0.14	-0.75	-0.53	0.24	0.48	0.28	-0.29	-0.04	-0.44
640	0.18	0	0.11	-0.1	-0.62	-0.07	-0.09	-0.09	-0.13	-0.59	-0.77	0.09	-0.61	-0.75	0.25	0.28	0.45	-0.7	-0.27	-0.44
641	-0.51	-0.26	-0.79	-0.64	0.82	-0.83	-0.66	-0.15	-0.7	-0.23	0.13	-0.92	0.04	0.25	-0.71	-0.29	-0.7	1.42	0	-0.17
642	-0.34	-0.31	0.12	0.09	0.4	-0.02	-0.11	-0.35	0.35	-0.35	-0.27	-0.12	-0.56	0.14	-0.2	-0.04	-0.27	0	0.84	-0.39
643	-0.06	-0.54	-0.69	-0.39	-0.19	-0.68	-0.42	-0.15	-0.59	0.41	0.41	-0.79	0.41	0.36	-0.47	-0.44	-0.44	-0.17	-0.39	0.54

Appendix C

CoEPrA Peptide Binding Affinity Data Sets

Publicly available peptide binding affinity data sets obtained from the literature are used in the experimental studies of this thesis. The peptide binding affinity data sets are obtained from a modeling competition [285]. Each task has a separate train (Table C.1) and test data set (Table C.2). A blind-validated experimental study conducted on these data sets. The columns correspond to peptide no, peptide residue, and expected real-value of binding affinity. The supplementary information of this thesis is accessible online at: `https://github.com/vuslan/pepbnd`.

TABLE C.1: List of peptides used to train the models of peptide binding affinity tasks.

Train Set of Task 1

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
1	ILDPFPVTD	2.94	46	IYDPFPVTV	5.41
2	ILDPFPVTY	3.19	47	YLSPGPVTA	5.44
3	ILDPFPVTH	3.60	48	LLFGYPVYV	5.45
4	SLHVGTQCA	3.79	49	YLFDGPVTA	5.50
5	HLLVGSSGL	3.91	50	ILDPFPVTT	5.54
6	NLQSLTNLL	3.96	51	RLWPLYPNV	5.57
7	SLNFMGYVI	4.00	52	YLFPGPVWA	5.59
8	ITSQVPFSV	4.06	53	YAIDLPVSV	5.63
9	VCMTVDSLV	4.20	54	YLFNGPVTV	5.65
10	LLMGTLGIV	4.21	55	ILDPFPVTF	5.67
11	ALIHHNTHL	4.30	56	YLWPGPVTV	5.70
12	MLDLQPETT	4.36	57	RLWPFYHNV	5.72
13	YVITTQHWL	4.39	58	YLAPGPVTA	5.74
14	ITFQVPFSV	4.42	59	IADPFPVTV	5.76
15	KTWGQYWQV	4.43	60	YLYPGPVTA	5.77
16	ITDQVPFSV	4.48	61	YLFPGPETA	5.81
17	LLAQFTSAI	4.51	62	ILDPFPVTP	5.82
18	VLHSFTDAI	4.54	63	FLWPFYPNV	5.89
19	ILDPFPVTK	4.59	64	FLDQVPFSV	5.98
20	YMNGTMSQV	4.67	65	FLWPFYHNV	5.99
21	ILDPFPVTW	4.71	66	ILWPLFHEV	6.03
22	FTDQVPFSV	4.76	67	ILWPLYPNV	6.06
23	KLHLYSHPI	4.77	68	ILDQVPFSV	6.09
24	ILDPFPVTS	4.78	69	ILNPFYPDV	6.11
25	YTDQVPFSV	4.80	70	FLWPLYPNV	6.14
26	IFDPFPVTV	4.89	71	FLNPFYPNV	6.16
27	CLTSTVQLV	4.93	72	FLNPIYHDV	6.16
28	YLWQYIFSV	4.94	73	YLFPGTVTA	6.16
29	IHDPFPVTV	4.96	74	YLCPGPVTA	6.18

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
30	RLMKQDFSV	4.97	75	YLFPPPVTV	6.19
31	VMGTLVALV	5.03	76	ILFPGPVTA	6.23
32	ILYQVPFSV	5.06	77	IIDPFPVTV	6.31
33	IPDPFPVTV	5.10	78	ILDPFPVTA	6.32
34	GLLGWSPQA	5.13	79	FLWPIYHNV	6.37
35	GLYSSTVPV	5.15	80	ILFPFVHSV	6.58
36	IISCTCPTV	5.17	81	ILDPFPVTG	6.66
37	FLCKQYLNL	5.21	82	YLFPFPITV	6.68
38	YLFPGPVTG	5.22	83	ILFPFPVEV	6.80
39	GTLGIVCPI	5.23	84	ILDDFPPTV	7.08
40	RLWPFYPNV	5.24	85	ILDPLPPTV	7.15
41	YLKPGPVTA	5.26	86	IMDPFPVTV	7.21
42	YLMPGPVTA	5.27	87	ILDPFPPPV	7.44
43	YMLDLQPET	5.28	88	ILDPFPITV	8.14
44	PLLPIFFCL	5.32	89	ILDPFPVTV	8.65
45	RLNPLYPNV	5.37			

Train Set of Task 2

No.	Peptide	Expected	No.	Peptide	Expected
1	FESTGNLD	5.010	39	FESTNNLI	7.748
2	FKSTGNLI	5.026	40	FDSTGNLI	7.814
3	FESTGNLR	5.232	41	FESTSNLI	7.821
4	FFSTGNLI	5.421	42	FESTWNLI	7.832
5	FESTGNLQ	5.687	43	FGSTGNLI	7.846
6	FESTGNLH	6.000	44	FESTGWLI	7.872
7	FESTGNLG	6.051	45	FESTINLI	7.887
8	FISTGNLI	6.329	46	FESDGNLI	7.890
9	QTFVVGCI	6.796	47	FESTLNLI	7.898
10	NEKSFKDI	6.910	48	FESTVNLI	7.912
11	FQSTGNLI	7.013	49	LEILNGEI	7.921
12	FLSTGNLI	7.088	50	FESTGKLI	7.927
13	FESTGNKI	7.159	51	DGLGGKLV	7.959

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
14	FESTGNLM	7.212	52	FESEGNLI	7.972
15	FESTGNDI	7.290	53	FESKGNLI	7.978
16	FESTGNLW	7.293	54	FEHTGNLN	7.982
17	KESTGNLI	7.308	55	FESWGNLI	7.989
18	FESTGNPI	7.410	56	FESTANLI	7.994
19	PESTGNLI	7.426	57	FEFTGNLN	8.000
20	FESTGNLA	7.455	58	FESTGVLI	8.023
21	FESTGNNI	7.521	59	FESAGNLI	8.031
22	FESTGNLS	7.525	60	FESPGNLI	8.042
23	FESTGNEI	7.541	61	FESTGNFI	8.044
24	VESTGNLI	7.545	62	FESTGNLI	8.046
25	FESTGNII	7.551	63	FESFGNLI	8.085
26	FESTGELI	7.593	64	FESRGNLI	8.095
27	HESTGNLI	7.607	65	FESYGNLI	8.099
28	FESTGNQI	7.612	66	FESTPNLI	8.141
29	AESTGNLI	7.624	67	FEATGNLN	8.178
30	SESTGNLI	7.641	68	FEDTGNLN	8.199
31	GESTGNLI	7.665	69	FEQTGNLN	8.217
32	FESTGDLI	7.683	70	FESTGRLI	8.222
33	IESTGNLI	7.715	71	FENTGNLN	8.224
34	MESTGNLI	7.716	72	FESVGNLI	8.230
35	QESTGNLI	7.727	73	FESIGNLI	8.239
36	NESTGNLI	7.736	74	FEGTGNLN	8.265
37	WESTGNLI	7.740	75	FERTGNLN	8.300
38	FESTGNHI	7.742	76	FELTGNLN	8.343

Train Set of Task 3

No.	Peptide	Expected	No.	Peptide	Expected
1	VVHFFKNIV	4.301	68	VLLDYQGML	7.095
2	VCMTVDSLV	5.146	69	LMIGTAAAV	7.102
3	LLGCAANWI	5.301	70	TVLRFVPPL	7.114
4	SAANDPIFV	5.342	71	NLGNLNVSI	7.119

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
5	TTAEEAAGI	5.380	72	ILHNGAYSL	7.127
6	LTVILGVLL	5.580	73	SIISAVVGI	7.159
7	LVSLLTFMI	5.716	74	VLAKDGTEV	7.174
8	QMTFHLFIA	5.778	75	YLEPGPVTI	7.187
9	ALPYWNFAT	5.820	76	FLYNRPLSV	7.212
10	FVTWHRYHL	5.869	77	FLWGPRALV	7.215
11	SLNFMGYVI	5.881	78	ILDQVPFSV	7.284
12	GIGILTVIL	6.000	79	ILSSLGLPV	7.301
13	IVMGNGTLV	6.001	80	LLFLGVVFL	7.301
14	SLSRFSWGA	6.041	81	YLVAYQATV	7.304
15	TVILGVLLL	6.072	82	YLEPGPVTV	7.342
16	WTDQVPFSV	6.145	83	ILSPFMPLL	7.347
17	AIAKAAAAV	6.176	84	YLSPGPVTA	7.383
18	ITSQVPFSV	6.196	85	IIDQVPFSV	7.398
19	ALAKAAAAI	6.211	86	YMNGTMSQV	7.398
20	GLGQVPLIV	6.301	87	FLCWGPFFL	7.415
21	LLSSNLSWL	6.342	88	LLFRFMRPL	7.447
22	SIIDPLIYA	6.342	89	ITWQVPFSV	7.457
23	YLVTRHADV	6.342	90	LLAVLYCLL	7.478
24	LIGNESFAL	6.380	91	GIRPYEILA	7.481
25	FLLPDAQSI	6.415	92	GLFLTTEAV	7.509
26	CLALSDLLV	6.447	93	YTYKWETFL	7.538
27	LLGRNSFEV	6.447	94	ALVGLFVLL	7.553
28	LLAVGATKV	6.477	95	SLDDYNHLV	7.583
29	MLLAVLYCL	6.478	96	FLLRWEQEI	7.592
30	AIYHPQQFV	6.504	97	SLLPAIVEL	7.620
31	ALAKAAAAL	6.511	98	YLSPGPVTV	7.642
32	FVNHRFTVV	6.523	99	GLIMVLSFL	7.658
33	WILRGTSFV	6.556	100	SLYADSPSV	7.658
34	TLDSQVMSL	6.580	101	RLLQETELV	7.682
35	GLYGAQYDV	6.602	102	IMDQVPFSV	7.719

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
36	MLASTLTDA	6.602	103	YLLPAIVHI	7.745
37	AIIDPLIYA	6.623	104	FLLLADARV	7.747
38	FLGGTPVCL	6.623	105	ALMDKSLHV	7.767
39	LMLPGMNGI	6.623	106	YLYPGPVTA	7.772
40	RLMIGTAAA	6.644	107	HMWNFISGI	7.818
41	LLFLLLADA	6.663	108	YLAPGPVTV	7.818
42	GTLGIVCPI	6.666	109	MLGTHTMEV	7.845
43	KLFPEVIDL	6.693	110	MTYAAPLFV	7.860
44	IAGGVMAVV	6.708	111	YLSQIAVLL	7.917
45	GLYRQWALA	6.733	112	YLMPGPVTV	7.932
46	MLQDMAILT	6.777	113	WLDQVPFSV	7.939
47	VILGVLLLI	6.785	114	SLYFGGICV	7.975
48	CLTSTVQLV	6.832	115	YLLALRYLA	8.000
49	ILLLCLIFL	6.845	116	SLLTFMIAA	8.027
50	DMWEHAFYL	6.879	117	GLMTAVYLV	8.051
51	ALTVVWLLV	6.893	118	FLLSLGIHL	8.053
52	LLPSLFLLL	6.903	119	FVVALIPLV	8.119
53	WMNRLIAFA	6.914	120	YLWPGPVTV	8.125
54	PLLPIFFCL	6.926	121	FLYGALRLA	8.149
55	ALAKAAAAA	6.947	122	LLLEAGALV	8.174
56	FLPWHRLFL	6.950	123	YLFPGPVTV	8.237
57	SLAGFVRML	6.954	124	ILFTFLHLA	8.268
58	TLGIVCPIC	6.964	125	RLPLVLPAV	8.292
59	KLTPLCVTL	6.991	126	YMDDVVLGV	8.301
60	LLCLIFLLV	6.996	127	GILTVILGV	8.342
61	RIWSWLLGA	7.000	128	NMVPFFPPV	8.403
62	SLLEIGEGV	7.009	129	FLYGAALLA	8.469
63	RLLDDTPEV	7.017	130	YLWPGPVTA	8.495
64	LLAGLVSLL	7.021	131	FLYGALALA	8.620
65	IAATYNFAV	7.032	132	FLDQVPFSV	8.658
66	YTDQVPFSV	7.066	133	ILWQVPFSV	8.770

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
67	SVMDPLIYA	7.079			

Train Set of Task 4

No.	Peptide	Expected	No.	Peptide	Expected
1	VVHFFKNIV	4.301	68	VLLDYQGML	7.095
2	VCMTVDSLV	5.146	69	LMIGTAAAV	7.102
3	LLGCAANWI	5.301	70	TVLRFVPPL	7.114
4	SAANDPIFV	5.342	71	NLGNLNVSI	7.119
5	TTAEEAAGI	5.380	72	ILHNGAYSL	7.127
6	LTVILGVLL	5.580	73	SIISAVVGI	7.159
7	LVSLLTFMI	5.716	74	VLAKDGTEV	7.174
8	QMTFHLFIA	5.778	75	YLEPGPVTI	7.187
9	ALPYWNFAT	5.820	76	FLYNRPLSV	7.212
10	FVTWHRYHL	5.869	77	FLWGPRALV	7.215
11	SLNFMGYVI	5.881	78	ILDQVPFSV	7.284
12	GIGILTVIL	6.000	79	ILSSLGLPV	7.301
13	IVMGNGTLV	6.001	80	LLFLGVVFL	7.301
14	SLSRFSWGA	6.041	81	YLVAYQATV	7.304
15	TVILGVLLL	6.072	82	YLEPGPVTV	7.342
16	WTDQVPFSV	6.145	83	ILSPFMPLL	7.347
17	AIAKAAAAV	6.176	84	YLSPGPVTA	7.383
18	ITSQVPFSV	6.196	85	IIDQVPFSV	7.398
19	ALAKAAAAI	6.211	86	YMNGTMSQV	7.398
20	GLGQVPLIV	6.301	87	FLCWGPFFL	7.415
21	LLSSNLSWL	6.342	88	LLFRFMRPL	7.447
22	SIIDPLIYA	6.342	89	ITWQVPFSV	7.457
23	YLVTRHADV	6.342	90	LLAVLYCLL	7.478
24	LIGNESFAL	6.380	91	GIRPYEILA	7.481
25	FLLPDAQSI	6.415	92	GLFLTTEAV	7.509
26	CLALSDLLV	6.447	93	YTYKWETFL	7.538
27	LLGRNSFEV	6.447	94	ALVGLFVLL	7.553
28	LLAVGATKV	6.477	95	SLDDYNHLV	7.583

Continued on next page

Table C.1 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
29	MLLAVLYCL	6.478	96	FLLRWEQEI	7.592
30	AIYHPQQFV	6.504	97	SLLPAIVEL	7.620
31	ALAKAAAAL	6.511	98	YLSPGPVTV	7.642
32	FVNHRFTVV	6.523	99	GLIMVLSFL	7.658
33	WILRGTSFV	6.556	100	SLYADSPSV	7.658
34	TLDSQVMSL	6.580	101	RLLQETELV	7.682
35	GLYGAQYDV	6.602	102	IMDQVPFSV	7.719
36	MLASTLTDA	6.602	103	YLLPAIVHI	7.745
37	AIIDPLIYA	6.623	104	FLLLADARV	7.747
38	FLGGTPVCL	6.623	105	ALMDKSLHV	7.767
39	LMLPGMNGI	6.623	106	YLYPGPVTA	7.772
40	RLMIGTAAA	6.644	107	HMWNFISGI	7.818
41	LLFLLLADA	6.663	108	YLAPGPVTV	7.818
42	GTLGIVCPI	6.666	109	MLGTHTMEV	7.845
43	KLFPEVIDL	6.693	110	MTYAAPLFV	7.860
44	IAGGVMAVV	6.708	111	YLSQIAVLL	7.917
45	GLYRQWALA	6.733	112	YLMPGPVTV	7.932
46	MLQDMAILT	6.777	113	WLDQVPFSV	7.939
47	VILGVLLLI	6.785	114	SLYFGGICV	7.975
48	CLTSTVQLV	6.832	115	YLLALRYLA	8.000
49	ILLLCLIFL	6.845	116	SLLTFMIAA	8.027
50	DMWEHAFYL	6.879	117	GLMTAVYLV	8.051
51	ALTVVWLLV	6.893	118	FLLSLGIHL	8.053
52	LLPSLFLLL	6.903	119	FVVALIPLV	8.119
53	WMNRLIAFA	6.914	120	YLWPGPVTV	8.125
54	PLLPIFFCL	6.926	121	FLYGALRLA	8.149
55	ALAKAAAAA	6.947	122	LLLEAGALV	8.174
56	FLPWHRLFL	6.950	123	YLFPGPVTV	8.237
57	SLAGFVRML	6.954	124	ILFTFLHLA	8.268
58	TLGIVCPIC	6.964	125	RLPLVLPAV	8.292
59	KLTPLCVTL	6.991	126	YMDDVVLGV	8.301

Continued on next page

Table C.1 – *Continued from previous page*

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
60	LLCLIFLLV	6.996	127	GILTVILGV	8.342
61	RIWSWLLGA	7.000	128	NMVPFFPPV	8.403
62	SLLEIGEGV	7.009	129	FLYGAALLA	8.469
63	RLLDDTPEV	7.017	130	YLWPGPVTA	8.495
64	LLAGLVSLL	7.021	131	FLYGALALA	8.620
65	IAATYNFAV	7.032	132	FLDQVPFSV	8.658
66	YTDQVPFSV	7.066	133	ILWQVPFSV	8.770
67	SVMDPLIYA	7.079			

TABLE C.2: List of peptides used to test the models of peptide binding affinity tasks.

Test Set of Task 1

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
1	YLFNGPVTA	5.80	45	IWDPFPVTV	5.13
2	IMDQVPFSV	5.71	46	YLFPGPSTA	5.69
3	RLLQETELV	4.83	47	KIFGSLAFL	4.40
4	HLESLFTAV	3.79	48	YLFPDPVTA	6.09
5	ILDPFPPTV	8.17	49	TLHEYMLDL	4.94
6	ILDPFPVTL	7.03	50	GILTVILGV	4.57
7	FLLSLGIHL	5.17	51	YLFPPPVTA	5.75
8	LQTTIHDII	3.90	52	RLWPIYHDV	5.55
9	IQDPFPVTV	6.05	53	SLDDYNHLV	5.27
10	VLLDYQGML	4.52	54	LLWFHISCL	4.13
11	FLWPIYHDV	6.16	55	VLIQRNPQL	5.06
12	TLGIVCPIC	4.68	56	YLFPGPMTA	5.98
13	YLFPGPVQA	6.14	57	HLYSHPIIL	5.41
14	FVTWHRYHL	4.21	58	WILRGTSFV	4.06
15	FLFPLPPEV	6.53	59	ILDPIPPTV	7.30
16	YLFPGPVTA	6.31	60	VTWHRYHLL	4.38
17	NLSWLSLDV	4.75	61	YLFPCPVTA	6.63
18	YLAPGPVTV	6.00	62	FLLTRILTI	4.95
19	ALPYWNFAT	4.66	63	IGDPFPVTV	3.92
20	ILDPFPVTE	3.13	64	MLGTHTMEV	5.37
21	ILDPFPVTQ	5.28	65	YLFPGVVTA	6.17
22	IDDPFPVTV	4.36	66	ILDPFPVTI	6.69
23	GLGQVPLIV	4.76	67	ILWPIYHNV	6.24
24	ALMPLYACI	5.08	68	YLEPGPVTL	5.41
25	GLSRYVARL	4.78	69	YLFPGPFTA	5.65
26	ILDDLPPTV	7.14	70	KLPQLCTEL	4.50
27	ILNPFYHNV	6.16	71	ILDPFPVTN	5.29
28	YLFDGPVTV	4.96	72	YLWDHFIEV	6.36
29	YLFQGPVTA	5.21	73	YLWQYIPSV	5.17

Continued on next page

Table C.2 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
30	SLYADSPSV	5.24	74	ILKEPVHGV	5.59
31	YLNPGPVTA	5.53	75	ILKPLYHNV	5.25
32	RLWPIYHNV	5.77	76	ITAQVPFSV	4.43
33	RLNPFYHDV	4.24	77	YLFPGPFTV	5.81
34	FLKPFYHNV	5.73	78	YLFPGPMTV	5.85
35	ILDPFPVTM	6.13	79	TTAEEAAGI	3.39
36	IVDPFPVTV	6.21	80	FLFPGPVTA	6.18
37	LMAVVLASL	3.99	81	WLDQVPFSV	5.23
38	ITDPFPVTV	6.08	82	FLDDHFCTV	6.68
39	ILWQVPFSV	5.91	83	SVYDFFVWL	5.12
40	ITWQVPFSV	5.01	84	ILDPFPVTC	5.65
41	ICDPFPVTV	5.45	85	ILDPFPPEV	7.68
42	ALCRWGLLL	4.91	86	NMVPFFPPV	5.60
43	ILDDFPVTV	7.16	87	ISDPFPVTV	5.50
44	SIISAVVGI	4.47	88	INDPFPVTV	4.78

Test Set of Task 2

1	YESTGNLI	7.740	39	FESTGHLI	7.997
2	FESTRNLI	7.679	40	FYSTGNLI	5.592
3	FESTGFLI	8.267	41	FPSTGNLI	8.113
4	FESTGTLI	7.922	42	DESTGNLI	7.712
5	FESTQNLI	7.819	43	FESQGNLI	8.094
6	FEKTGNLN	7.904	44	FESTKNLI	7.304
7	FEWTGNLN	8.225	45	FESTGNLL	7.737
8	FESTGQLI	7.920	46	FEVTGNLN	8.223
9	FASTGNLI	7.429	47	FLHPSMPV	7.149
10	FMSTGNLI	6.863	48	FESTMNLI	7.888
11	FESLGNLI	8.403	49	FEITGNLN	8.197
12	FNSTGNLI	6.244	50	FWSTGNLI	5.325
13	FESTGNSI	7.612	51	FEPTGNLN	8.043
14	RESTGNLI	7.544	52	FESTGNLN	7.000
15	FESTGPLI	8.302	53	FHSTGNLI	5.122

Continued on next page

Table C.2 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
16	FESTDNLI	7.743	54	FEETGNLN	8.028
17	FESTGGLI	7.946	55	TESTGNLI	7.535
18	FTSTGNLI	7.547	56	FESTGNLK	5.010
19	FESTGNLT	7.293	57	FESTGSLI	7.992
20	FESTGNWI	7.974	58	FAFWAFVV	7.523
21	FESTGNLF	7.848	59	FESTGNRI	8.004
22	EESTGNLI	7.732	60	FESTGALI	7.964
23	FESTYNLI	7.460	61	LESTGNLI	7.716
24	FESTGNLP	5.919	62	FEYTGNLN	8.176
25	FESTGNGI	7.209	63	FEMTGNLN	8.222
26	FESTGILI	8.098	64	FESTGYLI	8.215
27	FESTGNVI	7.421	65	HAIHGLLV	7.319
28	FESTGMLI	7.979	66	FESTTNLI	7.821
29	FETTGNLN	8.232	67	FESTENLI	7.583
30	FESSGNLI	8.046	68	FAFPGELL	7.022
31	FESTGNLY	6.010	69	FESTGNLV	7.626
32	FESTHNLI	7.836	70	FESTGNYI	7.793
33	FESTGNTI	7.652	71	FESMGNLI	8.040
34	FESTGNAI	7.602	72	FESTGNMI	7.612
35	FVSTGNLI	7.216	73	FESHGNLI	8.248
36	FESTFNLI	7.895	74	FESTGLLI	8.079
37	FESNGNLI	7.880	75	FESGGNLI	7.985
38	AESKSVII	6.648	76	FSSTGNLI	7.718

Test Set of Task 3

1	GLYSSTVPV	7.577	68	AMVGAVLTA	7.122
2	FTDQVPFSV	7.212	69	ITAQVPFSV	7.020
3	VLIQRNPQL	7.644	70	ILLSIARVV	6.342
4	LLWFHISCL	6.682	71	FLYGALLAA	8.201
5	FMGAGSKAV	6.200	72	ALMPLYACI	8.000
6	FVWLHYYSV	7.821	73	GLYYLTTEV	7.682
7	ALAKAAAAM	7.398	74	GLLGWSPQA	8.027

Continued on next page

Table C.2 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
8	LLLCLIFLL	7.585	75	LLWQDPVPA	7.343
9	YAIDLPVSV	7.801	76	MLGNAPSVV	6.644
10	GLSRYVARL	7.174	77	SLADTNSLA	6.342
11	QVMSLHNLV	6.025	78	HLYSHPIIL	7.131
12	MMWYWGPSL	7.921	79	ALVLLMLPV	7.506
13	YLFPGPVTA	8.495	80	RMPAVTDLV	6.903
14	VLLPSLFLL	7.444	81	LLWSFQTSA	7.818
15	KIFGSLAFL	7.478	82	YLEPGPVTL	7.058
16	AVIGALLAV	7.747	83	ALAKAAAAV	6.597
17	ALLAGLVSL	7.117	84	YMLDLQPET	7.373
18	ALSTGLIHL	6.505	85	HLAVIGALL	6.986
19	YALTVVWLL	6.924	86	AMKADIQHV	6.777
20	YLDQVPFSV	8.638	87	RMFAANLGV	7.447
21	YVITTQHWL	6.877	88	IVGAETFYV	8.456
22	FLLTRILTI	8.073	89	LQTTIHDII	5.501
23	YMIMVKCWM	6.663	90	KLAGGVAVI	6.447
24	RLMKQDFSV	7.338	91	LLPLGYPFV	6.477
25	FLAGALLLA	6.223	92	ITFQVPFSV	7.179
26	FLEPGPVTA	6.898	93	GLYLSQIAV	7.017
27	LLAQFTSAI	7.301	94	LLVFACSAV	6.342
28	AVAKAAAAV	6.495	95	AMLQDMAIL	7.009
29	GLCFFGVAL	5.380	96	ILAGYGAGV	6.937
30	VIHAFQYVI	5.914	97	YLAPGPVTA	8.032
31	ILYQVPFSV	8.310	98	SLHVGTQCA	5.842
32	DLMGYIPLV	7.097	99	ILAQVPFSV	7.939
33	NLQSLTNLL	6.000	100	YLVSFGVWI	8.721
34	SVYVDAKLV	6.991	101	ALYGALLLA	8.143
35	RLLGSLNST	6.778	102	GLQDCTMLV	7.638
36	WLLIDTSNA	6.447	103	VLTALLAGL	7.086
37	KTWGQYWQV	7.957	104	FLYGALVLA	7.409
38	FLYGGLLLA	8.959	105	VLHSFTDAI	6.170

Continued on next page

Table C.2 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
39	ITDQVPFSV	6.947	106	ILTVILGVL	6.419
40	FAFRDLCIV	6.963	107	ITMQVPFSV	7.398
41	YLYPGPVTV	8.051	108	LLFGYPVYV	7.886
42	WLSLLVPFV	8.164	109	HLESLFTAV	5.301
43	TLLVVMGTL	5.580	110	RLTEELNTI	6.060
44	LLDVPTAAV	7.770	111	VMGTLVALV	7.547
45	YLYVHSPAL	8.268	112	SVYDFFVWL	7.289
46	AMFQDPQER	5.740	113	YLMPGPVTA	8.367
47	VVLGVVFGI	7.845	114	ITYQVPFSV	7.480
48	MALLRLPLV	7.279	115	ILSQVPFSV	7.699
49	HLYQGCQVV	6.832	116	RLVSGLVGA	6.818
50	IISCTCPTV	6.580	117	LLLLGLWGL	7.658
51	DPKVKQWPL	6.176	118	NLYVSLLLL	7.114
52	QLFEDNYAL	7.764	119	RMYGVLPWI	7.538
53	LMAVVLASL	6.954	120	FVNHDFTVV	6.523
54	LLSCLGCKI	5.342	121	ALIHHNTHL	6.623
55	VVMGTLVAL	7.069	122	ALCRWGLLL	7.000
56	VALVGLFVL	5.079	123	GLVDFVKHI	6.663
57	LLACAVIHA	6.602	124	ILDEAYVMA	6.623
58	VLAGLLGNV	7.721	125	GLLGNVSTV	7.620
59	YLSEGDMAA	6.532	126	HLLVGSSGL	5.792
60	KILSVFFLA	8.301	127	ILMQVPFSV	8.125
61	IMPGQEAGL	7.188	128	VLVGGVLAA	6.732
62	FLYGALLLA	8.585	129	AAAKAAAAV	6.398
63	ALLSDWLPA	7.025	130	VLLLDVTPL	7.301
64	GLACHQLCA	6.380	131	YLDLALMSV	8.260
65	YMDDVVLGA	6.699	132	WLEPGPVTA	6.082
66	QLFHLCLII	6.886	133	LLVVMGTLV	5.869
67	FVDYNFTIV	6.620			

Test Set of Task 4

1	RMFPNAPYL	91	25	IITEFMTYG	18

Continued on next page

Table C.2 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
2	YMFPNAPYL	110	26	IIIEFMTYG	46
3	SLGEQQYSV	104	27	IIIEFMTYV	80
4	YLGEQQYSV	89	28	KLGGGQYGE	17
5	ALLPAVPSL	116	29	KLGGGQYGV	42
6	YLLPAVPSL	100	30	YLGGGQFGV	111
7	NLGATLKGV	37	31	KLGGGQFGV	59
8	YLGATLKGV	64	32	YLINKEEAL	114
9	DLNALLPAV	15	33	KLLQRPVAV	58
10	YLNALLPAV	78	34	YLKALQRPV	63
11	GVFRGIQDV	24	35	VLNYGVCVC	18
12	GLRRGIQDV	22	36	VLNYGVCFC	18
13	KRYFKLSHL	27	37	VLWYGVCFC	63
14	KLYFKLSHL	93	38	VLNYGVCFV	90
15	ALLLRTPYS	25	39	VLWYGVCFV	121
16	ALLLRTPYV	94	40	VCGDENILV	46
17	CMTWNQMNL	85	41	FCGDENILV	41
18	YMTWNQMNL	67	42	FMGDENILV	74
19	EVYEGVWKK	16	43	FLGDENILV	87
20	KVYEGVWKK	18	44	QQNPSYDSV	17
21	KVYEGVWKV	70	45	FLNPSYDSV	89
22	KLGGGQYGV	42	46	KLNPSYDSV	58
23	KLGGGQYGV	42	47	YLNPSYDSV	83
24	YLGGGQYGV	78			

Appendix D

Mouse Class I MHC Alleles

Publicly available peptide binding affinity data sets obtained from the literature are used in the experimental studies of this thesis. Three mouse class I MHC peptide binding affinity data sets are obtained from a data set paper [286]. Mouse class I MHC peptide alleles, H2-Db, H2-Kb and H2-Kk are given in Table D.1, Table D.2 and Table D.3, respectively. A cross-validated experimental study conducted on these data sets. The columns correspond to peptide no, peptide residue, and expected real-value of binding affinity. The supplementary information of this thesis is accessible online at: `https://github.com/vuslan/pepbnd`.

TABLE D.1: List of epitopes used in cross-validated real-value binding affinity prediction of the H2-Db mouse class I MHC allele.

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
1	AAAENAEAA	7.357	34	RSVINIVII	5.854
2	AEDTNVSLI	3.357	35	SAIENLEYM	7.721
3	AENENMRTM	5.712	36	SEVSNVQRI	5.797
4	AMIENLEYM	7.620	37	SFYRNLLWL	6.542
5	ASNENIDTM	8.699	38	SGVENPGGY	4.881
6	ASNENMETM	7.750	39	SLLGNATAL	6.796
7	ASNENMRTM	8.155	40	SLLYNLDLM	8.097
8	CDFNNGITI	5.344	41	SMAENLEYM	7.222
9	CKGVNKEYL	7.409	42	SMIANLEYM	6.848
10	FAPGNYPAL	8.091	43	SMIEALEYM	6.796
11	FCGVNSDTV	6.799	44	SMIENAEYM	7.523
12	FQLCNSYDL	7.886	45	SMIENLAYM	6.780
13	FQPQNGQFI	8.067	46	SMIENLEAM	7.699
14	FRGPNVVTL	5.925	47	SMIENLEYA	7.538
15	GFKSNFNKI	3.357	48	SMIENLEYM	7.871
16	IISHNFCNL	6.027	49	SSVIGVWYL	5.854
17	IKPSNSEDL	5.538	50	SSVVGVWYL	6.268
18	ISANNDSEI	6.056	51	SSVVNVWYL	7.244
19	ISNGNSDCL	6.503	52	TAGANPMDL	4.658
20	ISVSNPGDL	6.658	53	TALANTIEV	8.444
21	ITYKNSTWV	6.570	54	TGICNQNII	7.699
22	KAVYNFATC	6.484	55	TGKLNLENL	4.754
23	KICQNFILL	5.606	56	VENPGGYCL	4.475
24	LIDYNKAAL	5.714	57	VKYPNLNDL	5.878
25	LLVFNYPGI	5.287	58	VLSFNLGDM	4.202
26	LTFTNDSII	5.835	59	VLSTNGDTL	6.370
27	LTFTNDSSI	5.824	60	WLVTNGSYL	6.911
28	NGLWNLDVI	8.000	61	YAIENAEAL	7.658
29	QAPTNRWML	8.252	62	YAIENAKAL	6.959

Continued on next page

Table D.1 – *Continued from previous page*

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
30	QGINNLDNL	7.824	63	YAIKNAEAL	7.678
31	QLPPNSLLI	3.533	64	YASDNQAIL	6.319
32	RGVINIVII	5.692	65	YSQGNSGLM	6.051
33	RLIQNSLTI	6.967			

TABLE D.2: List of epitopes used in cross-validated real-value binding affinity prediction of the H2-Kb mouse class I MHC allele.

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
1	RGYVYQGL	8.137	32	MWYWGPSL	5.125
2	SIINFEKL	8.138	33	VLLDYQGM	5.477
3	APGNYPAL	6.558	34	YSILSPFL	5.954
4	FSVIFDRL	6.971	35	ANEGYDAL	4.924
5	IGRFYIQM	7.770	36	DDEEYVIL	3.907
6	KSSFYRNL	7.066	37	GTYHFTKL	7.745
7	KVVRFDKL	7.310	38	HDQLFSLL	5.639
8	LSYSAGAL	7.523	39	HPTLFKVL	6.208
9	MGLIYNRM	8.337	40	HPYLYRLL	6.712
10	MITQFESL	7.398	41	ISFAFCQL	8.886
11	MMIWHSNL	6.564	42	LIFNYPGV	7.398
12	MNIQFTAV	7.602	43	LIYNYPGV	8.387
13	MNYYWTLL	7.284	44	LMSGFRQM	5.162
14	RFYRTCKL	7.377	45	LQQRYSRL	9.222
15	RGYVFQGL	8.509	46	LVYNYPGV	7.638
16	RSYLIRAL	7.174	47	NHPVFSPL	7.252
17	RTFSFQNI	8.013	48	NTVVFDAL	3.810
18	SSIEFARL	8.770	49	QESCYGRL	6.463
19	SSISFCGV	8.678	50	QPQNYLRL	4.287
20	SSLPFQNI	8.056	51	SIILFLPL	9.000
21	VYIEVLHL	7.699	52	SKLQYKII	3.810
22	VYINTALL	7.886	53	VDYNFTIV	7.444
23	AIIKFAAL	8.046	54	ALISFLLL	6.030
24	RGYKYQGL	7.854	55	GVYQFKSV	8.000
25	ASARFSWL	6.523	56	ISHNFCNL	6.431
26	CLIFLLVL	5.222	57	IVTMFEAL	7.174
27	FIIFLFIL	5.301	58	LVSIFLHL	5.553
28	FVQWFVGL	6.824	59	NSHHYISM	5.507
29	IIFLFILL	5.125	60	SQTSYQYL	5.729

Continued on next page

Table D.2 – *Continued from previous page*

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
30	ILSPFLPL	6.329	61	TSYQYLII	7.469
31	LSSIFSRI	5.477	62	YTVKYPNL	6.770

TABLE D.3: List of epitopes used in cross-validated real-value binding affinity prediction of the H2-Kk mouse class I MHC allele.

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
1	AESKSVII	6.648	78	FESTGNLY	6.010
2	NEKSFKDI	6.910	79	FESTGNMI	7.612
3	QTFVVGCI	6.796	80	FESTGNNI	7.521
4	AESTGNLI	7.624	81	FESTGNPI	7.410
5	DESTGNLI	7.712	82	FESTGNQI	7.612
6	EESTGNLI	7.732	83	FESTGNRI	8.004
7	FASTGNLI	7.429	84	FESTGNSI	7.612
8	FDSTGNLI	7.814	85	FESTGNTI	7.652
9	FEATGNLN	8.178	86	FESTGNVI	7.421
10	FEDTGNLN	8.199	87	FESTGNWI	7.974
11	FEETGNLN	8.028	88	FESTGNYI	7.793
12	FEFTGNLN	8.000	89	FESTGPLI	8.302
13	FEGTGNLN	8.265	90	FESTGQLI	7.920
14	FEHTGNLN	7.982	91	FESTGRLI	8.222
15	FEITGNLN	8.197	92	FESTGSLI	7.992
16	FEKTGNLN	7.904	93	FESTGTLI	7.922
17	FELTGNLN	8.343	94	FESTGVLI	8.023
18	FEMTGNLN	8.222	95	FESTGWLI	7.872
19	FENTGNLN	8.224	96	FESTGYLI	8.215
20	FEPTGNLN	8.043	97	FESTHNLI	7.836
21	FEQTGNLN	8.217	98	FESTINLI	7.887
22	FERTGNLN	8.300	99	FESTKNLI	7.304
23	FESAGNLI	8.031	100	FESTLNLI	7.898
24	FESDGNLI	7.890	101	FESTMNLI	7.888
25	FESEGNLI	7.972	102	FESTNNLI	7.748
26	FESFGNLI	8.085	103	FESTPNLI	8.141
27	FESGGNLI	7.985	104	FESTQNLI	7.819
28	FESHGNLI	8.248	105	FESTRNLI	7.679
29	FESIGNLI	8.239	106	FESTSNLI	7.821

Continued on next page

Table D.3 – *Continued from previous page*

No.	Peptide	Expected	No.	Peptide	Expected
30	FESKGNLI	7.978	107	FESTTNLI	7.821
31	FESLGNLI	8.403	108	FESTVNLI	7.912
32	FESMGNLI	8.040	109	FESTWNLI	7.832
33	FESNGNLI	7.880	110	FESTYNLI	7.460
34	FESPGNLI	8.042	111	FESVGNLI	8.230
35	FESQGNLI	8.094	112	FESWGNLI	7.989
36	FESRGNLI	8.095	113	FESYGNLI	8.099
37	FESSGNLI	8.046	114	FETTGNLN	8.232
38	FESTANLI	7.994	115	FEVTGNLN	8.223
39	FESTDNLI	7.743	116	FEWTGNLN	8.225
40	FESTENLI	7.583	117	FEYTGNLN	8.176
41	FESTFNLI	7.895	118	FFSTGNLI	5.421
42	FESTGALI	7.964	119	FGSTGNLI	7.846
43	FESTGDLI	7.683	120	FHSTGNLI	5.122
44	FESTGELI	7.593	121	FISTGNLI	6.329
45	FESTGFLI	8.267	122	FKSTGNLI	5.026
46	FESTGGLI	7.946	123	FLSTGNLI	7.088
47	FESTGHLI	7.997	124	FMSTGNLI	6.863
48	FESTGILI	8.098	125	FNSTGNLI	6.244
49	FESTGKLI	7.927	126	FPSTGNLI	8.113
50	FESTGLLI	8.079	127	FQSTGNLI	7.013
51	FESTGMLI	7.979	128	FRSTGNLI	4.192
52	FESTGNAI	7.602	129	FSSTGNLI	7.718
53	FESTGNDI	7.290	130	FTSTGNLI	7.547
54	FESTGNEI	7.541	131	FVSTGNLI	7.216
55	FESTGNFI	8.044	132	FWSTGNLI	5.325
56	FESTGNGI	7.209	133	FYSTGNLI	5.592
57	FESTGNHI	7.742	134	GESTGNLI	7.665
58	FESTGNII	7.551	135	HESTGNLI	7.607
59	FESTGNKI	7.159	136	IESTGNLI	7.715
60	FESTGNLA	7.455	137	KESTGNLI	7.308

Continued on next page

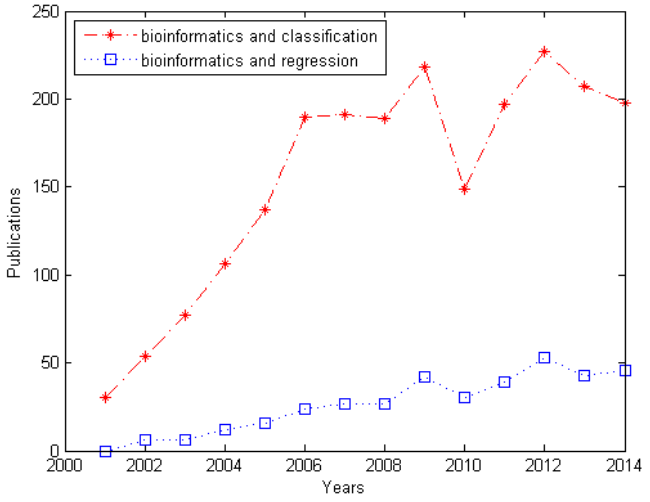
Table D.3 – *Continued from previous page*

No.	**Peptide**	**Expected**	**No.**	**Peptide**	**Expected**
61	FESTGNLD	5.010	138	LESTGNLI	7.716
62	FESTGNLE	4.707	139	MESTGNLI	7.716
63	FESTGNLF	7.848	140	NESTGNLI	7.736
64	FESTGNLG	6.051	141	PESTGNLI	7.426
65	FESTGNLH	6.000	142	QESTGNLI	7.727
66	FESTGNLI	8.046	143	RESTGNLI	7.544
67	FESTGNLK	5.010	144	SESTGNLI	7.641
68	FESTGNLL	7.737	145	TESTGNLI	7.535
69	FESTGNLM	7.212	146	VESTGNLI	7.545
70	FESTGNLN	7.000	147	WESTGNLI	7.740
71	FESTGNLP	5.919	148	YESTGNLI	7.740
72	FESTGNLQ	5.687	149	DGLGGKLV	7.959
73	FESTGNLR	5.232	150	FAFPGELL	7.022
74	FESTGNLS	7.525	151	FAFWAFVV	7.523
75	FESTGNLT	7.293	152	FLHPSMPV	7.149
76	FESTGNLV	7.626	153	HAIHGLLV	7.319
77	FESTGNLW	7.293	154	LEILNGEI	7.921

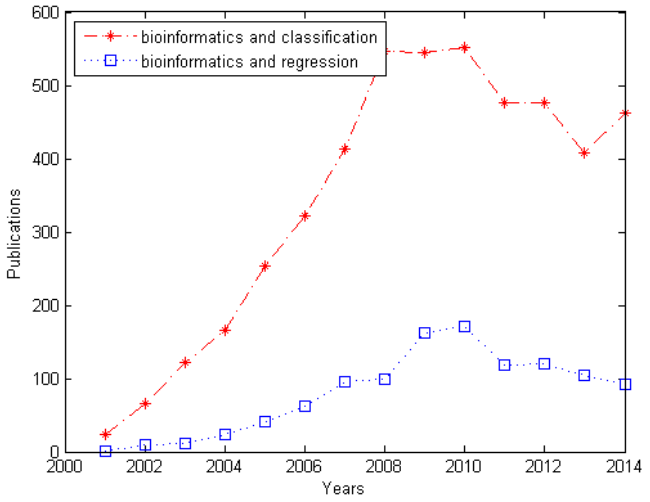
Appendix E

Graphs of the Keyword Sets

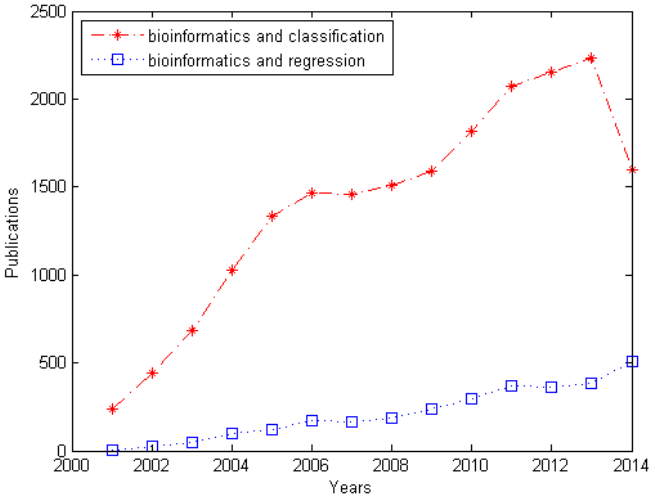
This appendix provides graphs related to the prediction studies in bioinformatics and systems biology. The keyword sets; "systems biology and regression", "bioinformatics and regression", "computational biology and prediction and regression", "systems biology and prediction and regression", "bioinformatics and prediction and regression" were used to reveal the papers from the well-known academic research databases such as Scopus, Web of Science, and PubMed. According to highly respected academic research databases, the number of publications per year in the fields of classification and regression are shown in Fig. E.1 - Fig. E.5.



(a) Web of Science

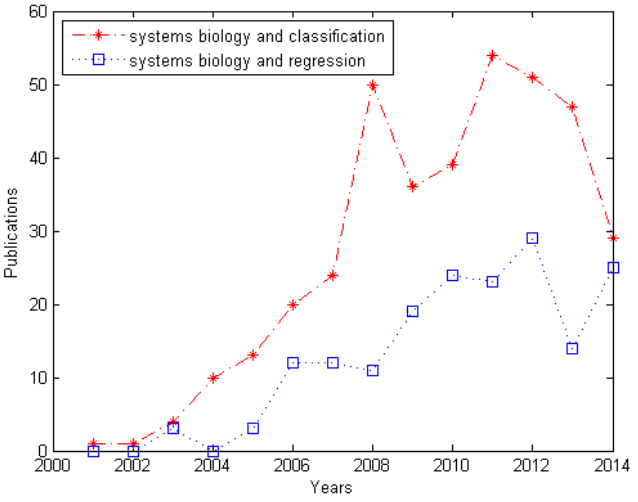


(b) Scopus

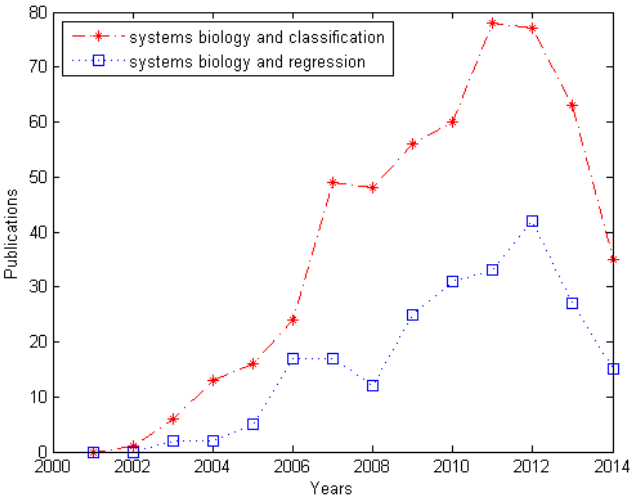


(c) PubMed

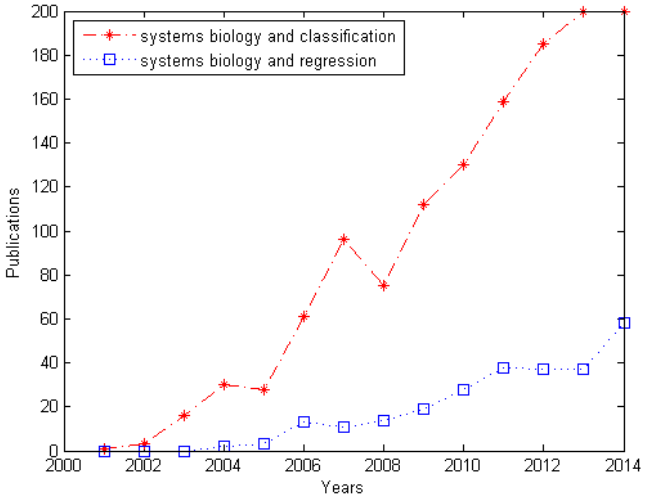
FIGURE E.1: Number of publications per year in respected databases related to the keywords: 1) bioinformatics and classification 2) bioinformatics and regression.



(a) Web of Science

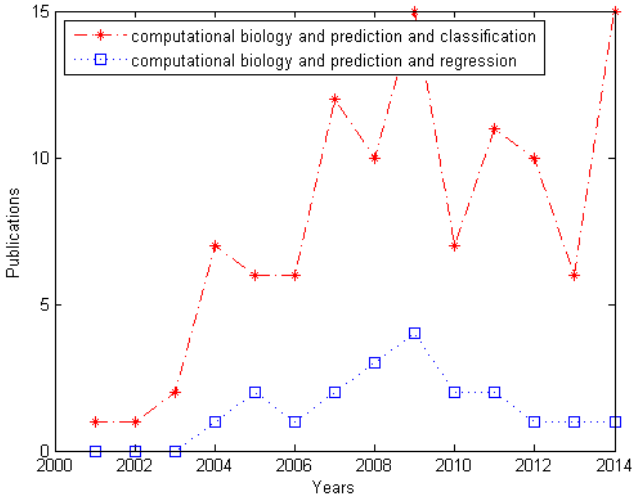


(b) Scopus

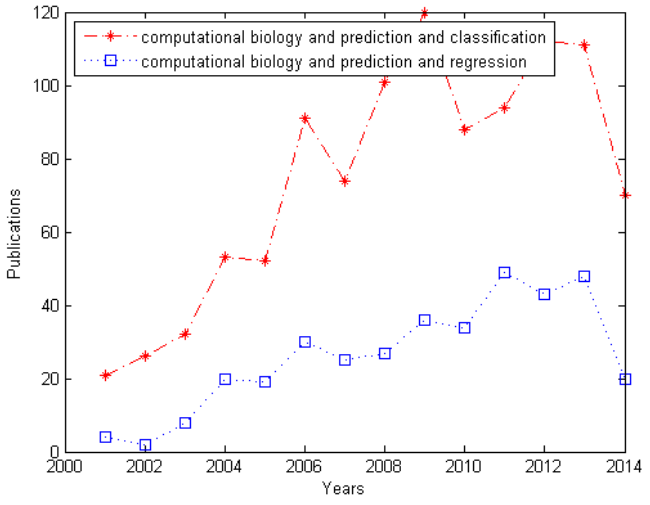


(c) PubMed

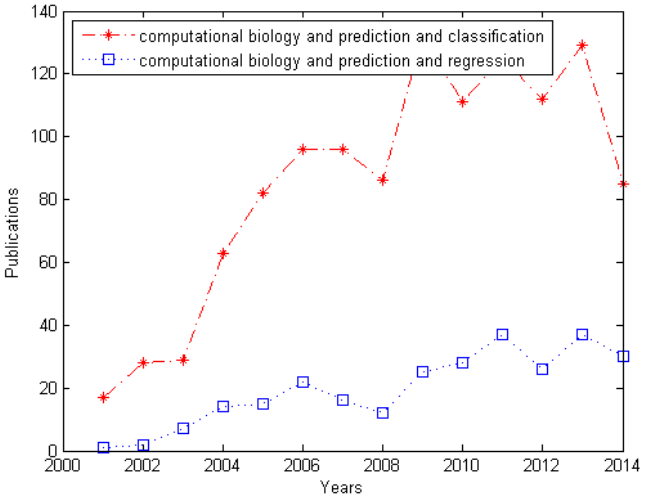
FIGURE E.2: Number of publications per year in respected databases related to the keywords: 1) systems biology and classification 2) systems biology and regression.



(a) Web of Science

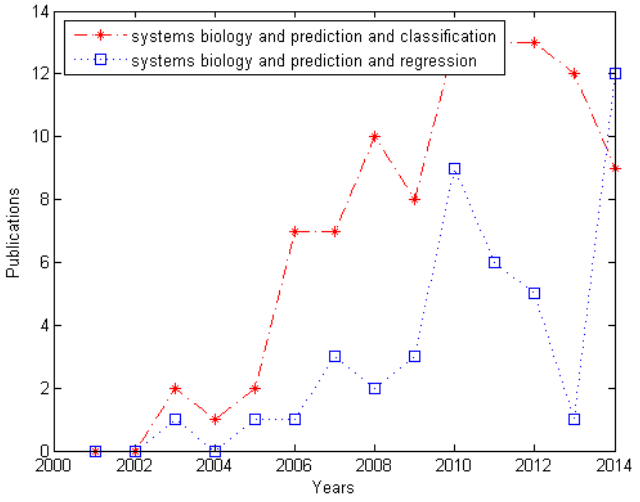


(b) Scopus

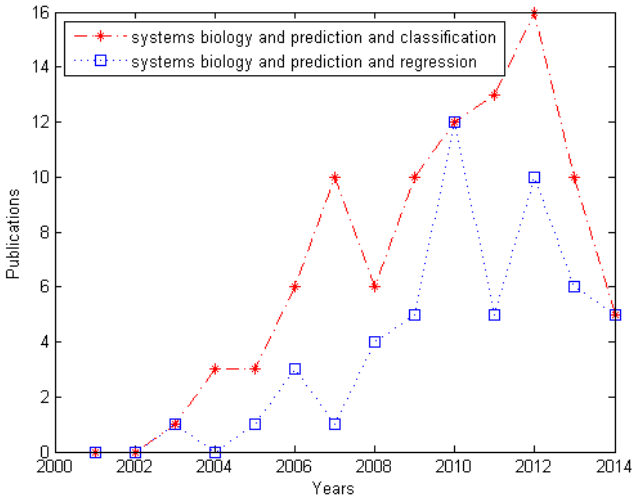


(c) PubMed

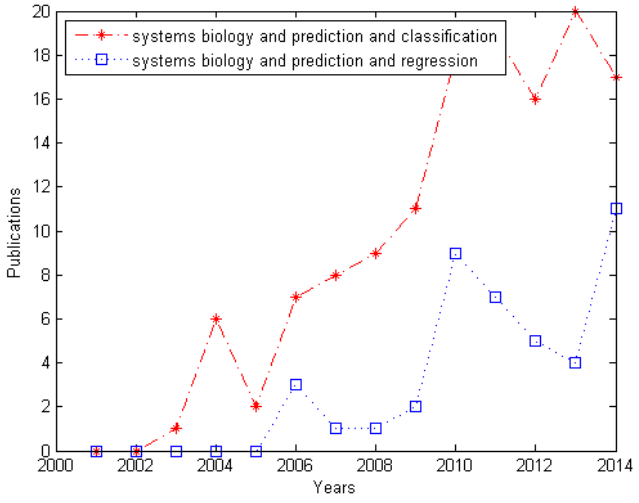
FIGURE E.3: Number of publications per year in respected databases related to the keywords: 1) computational biology and prediction and classification 2) computational biology and prediction and regression.



(a) Web of Science

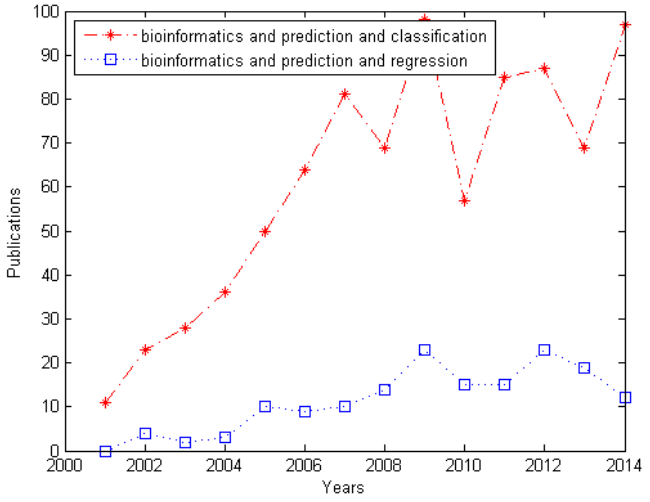


(b) Scopus

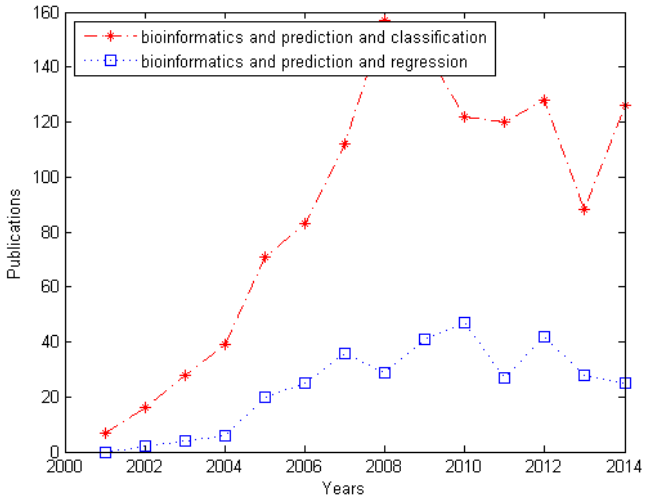


(c) PubMed

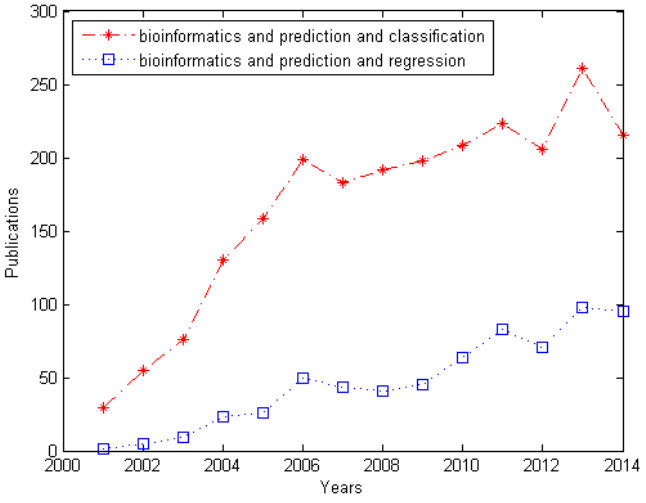
FIGURE E.4: Number of publications per year in respected databases related to the keywords: 1) systems biology and prediction and classification 2) systems biology and prediction and regression.



(a) Web of Science



(b) Scopus



(c) PubMed

FIGURE E.5: Number of publications per year in respected databases related to the keywords: 1) bioinformatics and prediction and classification 2) bioinformatics and prediction and regression.

References

[1] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, and Robert A. Holt. The sequence of the human genome. *Science*, 291(5507): 1304–1351, 2001.

[2] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, and William FitzHugh. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[3] Li Ding, Michael C. Wendl, Daniel C. Koboldt, and Elaine R. Mardis. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Human Molecular Genetics*, 19(R2):R188–R196, 2010.

[4] Christine Orengo, David T Jones, and Janet M Thornton. *Bioinformatics: Genes, Proteins and Computers*. Garland Science, 2003.

[5] Timothy J. Ross. *Fuzzy Logic with Engineering Applications*. John Wiley & Sons, 2004.

[6] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-15(1):116–132, 1985.

[7] M. Sugeno and G. T. Kang. Structure identification of fuzzy model. *Fuzzy Sets and Systems*, 28(1):15–33, 1988.

[8] O. Cordon, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena. Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*, 141(1):5–31, 2004.

[9] R. Nikhil, P. Kuhu, J. C. Bezdek, and T. A. Runkler. Some issues in system identification using clustering. In *International Conference on Neural Networks*, volume 4, pages 2524–2529, 1997.

[10] Magne Setnes. Supervised fuzzy clustering for rule extraction. *Fuzzy Systems, IEEE Transactions on*, 8(4):416–424, 2000.

[11] Joelle Presson and Jan Jenner. *Biology: dimensions of life*. McGraw Hill, 2008.

[12] Peter Lydyard, Alex Whelan, and Michael Fanger. *BIOS Instant Notes in Immunology*. Taylor & Francis, 2011.

[13] PJ Bjorkman, MA Saper, B. Samraoui, WS Bennett, JL Strominger, and DC Wiley. Structure of the human class I histocompatibility antigen, HLA-A2. *The Journal of Immunology*, 174(1):6–12, 2005.

[14] Paul AH Moss, William MC Rosenberg, and John I Bell. The human T cell receptor in health and disease. *Annual Review of Immunology*, 10(1):71–96, 1992.

[15] J. Andrew Bristol, Jeffrey Schlom, and Scott I. Abrams. Development of a murine mutant Ras CD8+ CTL peptide epitope variant that possesses enhanced MHC class I binding and immunogenic properties. *The Journal of Immunology*, 160(5): 2433–2441, 1998.

[16] A. Sette, S. Buus, E. Appella, J. A. Smith, R. Chesnut, C. Miles, S. M. Colon, and H. M. Grey. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci USA*, 86(9): 3296–3300, 1989.

[17] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152(1):163–175, 1994.

[18] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50 (3-4):213–219, 1999.

[19] P. A. Reche, J. P. Glutting, and E. L. Reinherz. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol*, 63(9):701–709, 2002.

[20] Pierre Dönnes and Arne Elofsson. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3(1):25, 2002.

[21] Jennifer D. Stone, Adam S. Chervin, and David M. Kranz. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology*, 126 (2):165–176, 2009.

[22] Kirsten Roomp, Iris Antes, and Thomas Lengauer. Predicting MHC class I epitopes in large datasets. *BMC Bioinformatics*, 11(1):90, 2010.

[23] Robert D. Bremel and E. Jane Homan. An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches. *Immunome Research*, 6(1): 7, 2010.

[24] Manoj Bhasin and G. P. S. Raghava. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Science*, 13(3):596–607, 2004.

[25] Wen Liu, Xiangshan Meng, Qiqi Xu, Darren Flower, and Tongbin Li. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, 7(1):182, 2006.

[26] Volkan Uslan and Huseyin Seker. Support vector-based Takagi-Sugeno fuzzy system for the prediction of binding affinity of peptides. In *Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE*, pages 4062–4065, 2013.

[27] Volkan Uslan and Huseyin Seker. Support vector-based fuzzy system for the prediction of mouse class I MHC peptide binding affinity. In *Bioinformatics and Bioengineering (BIBE), IEEE 13th International Conference on*, pages 1–4, 2013.

[28] Volkan Uslan and Huseyin Seker. Modelling non-linear system in the post genome era: Quantitative prediction of degree of peptide binding by using support vector based fuzzy system (in preparation).

[29] Volkan Uslan, Huseyin Seker, and Robert I. John. Modelling non-linear system in the post genome era: Quantitative prediction of degree of peptide binding by using support vector based type-2 fuzzy system (in preparation).

[30] Volkan Uslan, Huseyin Seker, and Robert I. John. A support vector-based interval type-2 fuzzy system. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 2396–2401, 2014.

[31] Volkan Uslan and Huseyin Seker. Survey on quantitative prediction in bioinformatics, systems and computational biology. (in preparation).

[32] Javier Herrero, Fatima Al-Shahrour, Ramon Diaz-Uriarte, Alvaro Mateos, Juan M Vaquerizas, Javier Santoyo, and Joaquin Dopazo. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Research*, 31(13): 3461–3467, 2003.

[33] Wei Guan, Manshui Zhou, Christina Y Hampton, Benedict B Benigno, L DeEtte Walker, Alexander Gray, John F McDonald, and Facundo M Fernández. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10(1):259, 2009.

[34] Walter P. Blackstock and Malcolm P. Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17(3):121–127, 1999.

[35] Pascal Roy, Caroline Truntzer, Delphine Maucort-Boulch, Thomas Jouve, and Nicolas Molinari. Protein mass spectra data analysis for clinical biomarker discovery: a global review. *Briefings in Bioinformatics*, 12(2):176–186, 2011.

[36] Ting Huang, Jingjing Wang, Weichuan Yu, and Zengyou He. Protein inference: a review. *Briefings in Bioinformatics*, 2012.

[37] J. Shi, B. Chen, and F. X. Wu. Unifying protein inference and peptide identification with feedback to update consistency between peptides. *Proteomics*, 13(2): 239–247, 2013.

[38] Ting Huang, Haipeng Gong, Can Yang, and Zengyou He. ProteinLasso: A Lasso regression approach to protein inference problem in shotgun proteomics. *Computational Biology and Chemistry*, 43:46–54, 2013.

[39] Steven Gay, Pierre-Alain Binz, Denis F. Hochstrasser, and Ron D. Appel. Peptide mass fingerprinting peak intensity prediction: Extracting knowledge from spectra. *Proteomics*, 2(10):1374–1391, 2002.

[40] Wiebke Timm, Sebastian Böcker, Thorsten Twellmann, and Tim Wilhelm Nattkemper. Peak intensity prediction for PMF mass spectra using support vector regression. In *Proc. of the 7th International FLINS Conference on Applied Artificial Intelligence*, pages 565–572.

[41] Wiebke Timm, Alexandra Scherbart, Sebastian Böcker, Oliver Kohlbacher, and Tim W Nattkemper. Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. *BMC Bioinformatics*, 9(1):443, 2008.

[42] Anuj R. Shah, Khushbu Agarwal, Erin S. Baker, Mudita Singhal, Anoop M. Mayampurath, Yehia M. Ibrahim, Lars J. Kangas, Matthew E. Monroe, Rui Zhao, Mikhail E. Belov, Gordon A. Anderson, and Richard D. Smith. Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bioinformatics*, 26(13):1601–1607, 2010.

[43] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[44] Alexandra Scherbart, Wiebke Timm, Sebastian Böcker, and Tim W Nattkemper. Neural network approach for mass spectrometry prediction by peptide prototyping. In *Artificial Neural Networks (ICANN)*, pages 90–99. Springer, 2007.

[45] M. Gorania, H. Seker, and P. I. Haris. Predicting a protein's melting temperature from its amino acid sequence. In *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, pages 1820–1823, 2010.

[46] Jing Yan, Marcin J Mizianty, Paul L Filipow, Vladimir N Uversky, and Lukasz Kurgan. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(8):1671–1680, 2013.

[47] Yi-Hung Huang, John P Rice, Scott F Saccone, José L Ambite, Yigal Arens, Jay A Tischfield, and Chun-Nan Hsu. A ν-support vector regression based approach for predicting imputation quality. In *BMC Proceedings*, volume 6, page S3, 2012.

[48] Yu Zhang, Xiuwen Liu, and Jonathan H. Dennis. Quantitative models for statistical nucleosome occupancy prediction. In *Bioinformatics and Biomedicine Workshops (BIBMW), IEEE International Conference on*, pages 937–939, 2012.

[49] H Tomas Rube and Jun S Song. Quantifying the role of steric constraints in nucleosome positioning. *Nucleic Acids Research*, 42(4):2147–2158, 2014.

[50] Ander Muniategui, Rubén Nogales-Cadenas, Miguél Vázquez, Xabier L Aranguren, Xabier Agirre, Aernout Luttun, Felipe Prosper, Alberto Pascual-Montano, and Angel Rubio. Quantification of miRNA-mRNA interactions. *PloS One*, 7(2):e30766, 2012.

[51] Qi Liu, Qian Xu, Vincent Zheng, Hong Xue, Zhiwei Cao, and Qiang Yang. Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study. *BMC Bioinformatics*, 11(1):181, 2010.

[52] Peng Jiang, Xiao Sun, and Zuhong Lu. Quantitative estimation of siRNAs gene silencing capability by random forest regression model. In *Bioinformatics and Biomedical Engineering (ICBBE), The 1st International Conference on*, pages 230–233, 2007.

[53] Fantine Mordelet, John Horton, Alexander J. Hartemink, Barbara E. Engelhardt, and Raluca Gordon. Stability selection for regression-based models of transcription factor DNA binding specificity. *Bioinformatics*, 29(13):i117–i125, 2013.

[54] Erdal Cosgun, Nita A. Limdi, and Christine W. Duarte. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in african americans. *Bioinformatics*, 27(10):1384–1389, 2011.

[55] Gustavo De Los Campos, Hugo Naya, Daniel Gianola, Jos Crossa, Andrs Legarra, Eduardo Manfredi, Kent Weigel, and Jos Miguel Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182:375–385, 2009.

[56] Y. Chen, X. Wu, and R. Jiang. Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, page 57, 2013.

[57] Tao Huang, Baolin Wu, Paul Lizardi, and Hongyu Zhao. Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, 21 (20):3811–3817, 2005.

[58] Paul HC Eilers and Renée X De Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153, 2005.

[59] Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[60] W. Deng, C. M. Breneman, and M. J. Embrechts. Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *Bioinformatics*, 44(2):699–703, 2004.

[61] Michael P Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, and Julio Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PloS One*, 8(4):e61318, 2013.

[62] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[63] Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[64] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems*, 96(1):86–103, 2009.

[65] Zeke S. H. Chan, Ilkka Havukkala, Vishal Jain, Yingjie Hu, and Nikola Kasabov. Soft computing methods to predict gene regulatory networks: An integrative approach on time-series gene expression data. *Applied Soft Computing*, 8(3):1189–1199, 2008.

[66] Jie Xiong and Tong Zhou. Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses. *PloS One*, 7(9): e43819, 2012.

[67] Xun Huang and Zhike Zi. Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR). *Molecular BioSystems*, 2014.

[68] Michael Andrec, Boris N. Kholodenko, Ronald M. Levy, and Eduardo Sontag. Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. *Journal of Theoretical Biology*, 232(3):427–441, 2005.

[69] Belhassen Bayar, Nidhal Bouaynaya, and Roman Shterenberg. Inference of genetic regulatory networks with unknown covariance structure. In *GENSiPS*, pages 74–77, 2013.

[70] Jing Qin, Yaohua Hu, Feng Xu, Hari Krishna Yalamanchili, and Junwen Wang. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 67(3):294–303, 2014.

[71] Zixing Wang, Wenlong Xu, F. Anthony San Lucas, and Yin Liu. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 2013.

[72] Jochen Supper, Holger Fröhlich, Christian Spieth, Andreas Dräger, and Andreas Zell. Inferring gene regulatory networks by machine learning methods. In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference (APBC)*, volume 5, pages 247–256, 2007.

[73] M. K. Yeung, J. Tegner, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6163–6168, 2002.

[74] Céline Brouard, Marie Szafranski, Florence D'Alché-Buc, et al. Protein-protein interaction network inference with semi-supervised Output Kernel Regression. *Proc. de la 13ème Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, pages 133–136, 2012.

[75] Ala Qabaja, Mohammed Alshalalfa, Tarek A Bismar, and Reda Alhajj. Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):1–11, 2013.

[76] Sara Berthoumieux, Matteo Brilli, Daniel Kahn, Hidde De Jong, and Eugenio Cinquemani. On the identifiability of metabolic network models. *Journal of Mathematical Biology*, 67(6-7):1795–1832, 2013.

[77] A. Castellini, M. Zucchelli, M. Busato, and V. Manca. From time series to biological network regulations: An evolutionary approach. *Molecular BioSystems*, 9(2): 225–233, 2013.

[78] Ji Wan, Wen Liu, Qiqi Xu, Yongliang Ren, Darren R. Flower, and Tongbin Li. SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics*, 7 (1):463, 2006.

[79] Irini A. Doytchinova and Darren R. Flower. Predicting class I major histocompatibility complex (MHC) binders using multivariate statistics: comparison of discriminant analysis and multiple linear regression. *Journal of Chemical Information and Modeling*, 47(1):234–238, 2007.

[80] Sébastien Giguère, Mario Marchand, François Laviolette, Alexandre Drouin, and Jacques Corbeil. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinformatics*, 14(1):82, 2013.

[81] Ozgur Demir-Kavuk, Mayumi Kamada, Tatsuya Akutsu, and Ernst-Walter Knapp. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinformatics*, 12(1):412, 2011.

[82] Ovidiu Ivanciuc and Werner Braun. Robust quantitative modeling of peptide binding affinities for MHC molecules using physical-chemical descriptors. *Protein and Peptide Letters*, 14(9):903, 2007.

[83] Channa K. Hattotuwagama, Christopher P. Toseland, Pingping Guan, Debra J. Taylor, Shelley L. Hemsley, I. A. Doytchinova, and Darren R. Flower. Toward prediction of class II mouse major histocompatibility complex peptide binding affinity: in silico bioinformatic evaluation using partial least squares, a robust multivariate statistical technique. *Journal of Chemical Information and Modeling*, 46(3):1491–1502, 2006.

[84] Linyuan Guo, Cheng Luo, and Shanfeng Zhu. MHC2SKpan: a novel kernel based approach for pan-specific MHC class II peptide binding prediction. *BMC Genomics*, 14:S11, 2013.

[85] Xiaojian Shao, Chris SH Tan, Courtney Voss, Shawn SC Li, Naiyang Deng, and Gary D Bader. A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain–peptide interaction from primary sequence. *Bioinformatics*, 27(3):383–390, 2011.

[86] I. A. Doytchinova and D. R. Flower. Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J Med Chem*, 44(22):3572–3581, 2001.

[87] I. A. Doytchinova and D. R. Flower. Quantitative approaches to computational vaccinology. *Immunology and Cell Biology*, 80(3):270–279, 2002.

[88] P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower. MHCPred: A server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*, 31(13): 3621–3624, 2003.

[89] Pingping Guan, Channa K Hattotuwagama, Irini A Doytchinova, and Darren R Flower. MHCPred 2.0: an updated quantitative T-cell epitope prediction server. *Applied bioinformatics*, 5(1):55–61, 2006.

[90] Andrew Bordner and Hans Mittelmann. Prediction of the binding affinities of peptides to class II MHC using a regularized thermodynamic model. *BMC Bioinformatics*, 11(1):41, 2010.

[91] Andrew Bordner and Hans Mittelmann. MultiRTA: A simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes. *BMC Bioinformatics*, 11(1):482, 2010.

[92] Stewart T. Chang, Debashis Ghosh, Denise E. Kirschner, and Jennifer J. Linderman. Peptide length-based prediction of peptide-MHC class II binding. *Bioinformatics*, 22(22):2761–2767, 2006.

[93] Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. Predicting MHC-II binding affinity using multiple instance regression. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(4):1067–1079, 2011.

[94] Xue-Ling Li, Min Zhu, Xiao-Lai Li, Hong-Qiang Wang, and Shulin Wang. Protein-Protein interaction affinity prediction based on interface descriptors and machine learning. In *Intelligent Computing Theories and Applications*, pages 205–212. Springer, 2012.

[95] Xueling Li, Min Zhu, Xiaolai Li, Hong-Qiang Wang, and Shulin Wang. Protein-protein binding affinity prediction based on an SVR ensemble. In *Intelligent Computing Technology*, pages 145–151. Springer, 2012.

[96] Yu Su, Ao Zhou, Xuefeng Xia, Wen Li, and Zhirong Sun. Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. *Protein Science*, 18(12):2550–2558, 2009.

[97] Christine Brun, Francois Chevenet, David Martin, Jerome Wojcik, Alain Guenoche, and Bernard Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6–R6, 2004.

[98] Albertha JM Walhout and Marc Vidal. Protein interaction maps for model organisms. *Nature Reviews Molecular Cell Biology*, 2(1):55–63, 2001.

[99] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[100] Matteo Pellegrini, David Haynor, and Jason M Johnson. Protein interaction networks. *Expert Review of Proteomics*, 1(2):239–249, 2004.

[101] Tero Aittokallio and Benno Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255, 2006.

[102] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1), 2007.

[103] John-Marc Chandonia and Marting Karplus. New methods for accurate prediction of protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*, 35(3):293–306, 1999.

[104] Xian-Ming Pan. Multiple linear regression for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 43(3):256–259, 2001.

[105] M. H. Zangooei and S. Jalili. Protein secondary structure prediction using DWKF based on SVR-NSGAII. *Neurocomputing*, 94:87–101, 2012.

[106] Murtada Khalafallah Elbashir, Yu Sheng, Jianxin Wang, FangXiang Wu, and Min Li. Predicting β-turns in protein using kernel logistic regression. *BioMed Research International*, 2013.

[107] Mehdi Poursheikhali Asgary, Samad Jahandideh, Parviz Abdolmaleki, and Anoshirvan Kazemnejad. Analysis and identification of β-turn types using multinomial logistic regression and artificial neural network. *Bioinformatics*, 23(23): 3125–3130, 2007.

[108] Burkhard Rost and Chris Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Bioinformatics*, 20 (3):216–226, 1994.

[109] Mohd Firdaus Raih, Shandar Ahmad, Rong Zheng, and Rahmah Mohamed. Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability. *Biophysical Chemistry*, 114(1):63 – 69, 2005.

[110] H S Chan and K A Dill. Origins of structure in globular proteins. *Proceedings of the National Academy of Sciences*, 87(16):6388–6392, 1990.

[111] Zheng Yuan, Kevin Burrage, and John S. Mattick. Prediction of protein solvent accessibility using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 48(3):566–570, 2002.

[112] Minh N. Nguyen and Jagath C. Rajapakse. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins: Structure, Function, and Bioinformatics*, 63(3):542–550, 2006.

[113] Zheng Yuan and Bixing Huang. Prediction of protein accessible surface areas by support vector regression. *Proteins: Structure, Function, and Bioinformatics*, 57 (3):558–564, 2004.

[114] Ke Chen, Michal Kurgan, and Lukasz Kurgan. Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. *Journal of Biomedical Science and Engineering*, 1(1):1–9, 2008.

[115] Darby TH Chang, Hsuan-Yu Huang, Yu-Tang Syu, and Chih-Peng Wu. Real value prediction of protein solvent accessibility using enhanced PSSM features. *BMC Bioinformatics*, 9(Suppl 12):S12, 2008.

[116] Shandar Ahmad, M. Michael Gromiha, and Akinori Sarai. Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 50(4):629–635, 2003.

[117] R. Adamczak, A. Porollo, and J. Meller. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins: Structure, Function, and Bioinformatics*, 56(4):753–767, 2004.

[118] Huiling Chen and Huan-Xiang Zhou. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Research*, 33(10), 2005.

[119] Jung-Ying Wang, Hahn-Ming Lee, and Shandar Ahmad. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins: Structure, Function, and Bioinformatics*, 61(3):481–491, 2005.

[120] Hongyi Zhou and Yaoqi Zhou. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins: Structure, Function, and Bioinformatics*, 54(2):315–322, 2004.

[121] Zhigang Xu, Chi Zhang, Song Liu, and Yaoqi Zhou. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins: Structure, Function, and Bioinformatics*, 63(4):961–966, 2006.

[122] Hua Zhang, Tuo Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. *BMC Bioinformatics*, 9(1):388, 2008.

[123] Jiangning Song, Hao Tan, Khalid Mahmood, Ruby HP Law, Ashley M Buckle, Geoffrey I Webb, Tatsuya Akutsu, and James C Whisstock. Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS One*, 4(9):e7072, 2009.

[124] Zheng Yuan and Zhi-Xin Wang. Quantifying the relationship of protein burying depth and sequence. *Proteins: Structure, Function, and Bioinformatics*, 70(2): 509–516, 2008.

[125] Liang-Tsung Huang and M. Michael Gromiha. First insight into the prediction of protein folding rate change upon point mutation. *Bioinformatics*, 26(17):2121–2127, 2010.

[126] Eshel Faraggi, Bin Xue, and Yaoqi Zhou. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins: Structure, Function, and Bioinformatics*, 74(4):847–856, 2009.

[127] Bin Xue, Ofer Dor, Eshel Faraggi, and Yaoqi Zhou. Real value prediction of backbone torsion angles. *Proteins: Structure, Function, and Bioinformatics*, 72 (1):427–433, 2008.

[128] Jiangning Song, Hao Tan, Mingjun Wang, Geoffrey I. Webb, and Tatsuya Akutsu. TANGLE: Two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. *PloS One*, 7(2):e30361, 2012.

[129] David Eramian, Narayanan Eswar, Min-Yi Shen, and Andrej Sali. How well can the accuracy of comparative protein structure models be predicted? *Protein Science*, 17(11):1881–1893, 2008.

[130] Jian Qiu, Will Sheffler, David Baker, and William Stafford Noble. Ranking predicted protein structures with support vector regression. *Proteins: Structure, Function, and Bioinformatics*, 71(3):1175–1182, 2008.

[131] Kristin Tøndel. Prediction of homology model quality with multivariate regression. *Journal of Chemical Information and Computer Sciences*, 44(5):1540–1551, 2004.

[132] Yifeng David Yang, Preston Spratt, Hao Chen, Changsoon Park, and Daisuke Kihara. Sub-AQUA: real-value quality assessment of protein structure models. *Protein Engineering Design and Selection*, 23(8):617–632, 2010.

[133] Gianluca Pollastri, Pierre Baldi, Pietro Fariselli, and Rita Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, 47(2):142–153, 2002.

[134] P Fariselli and R Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12(1):15–21, 1999.

[135] Jiangning Song and Kevin Burrage. Predicting residue-wise contact orders in proteins by support vector regression. *BMC Bioinformatics*, 7(1):425, 2006.

[136] Zheng Yuan. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, 6(1):248, 2005.

[137] Jiangning Song, Zheng Yuan, Hao Tan, Thomas Huber, and Kevin Burrage. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, 23(23):3147–3154, 2007.

[138] C. Savojardo, P. Fariselli, P. L. Martelli, and R. Casadio. Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations. *BMC Bioinformatics*, 14(1), 2013.

[139] Ole Lund, Kenneth Frimand, Jan Gorodkin, Henrik Bohr, Jakob Bohr, Jan Hansen, and Søren Brunak. Protein distance constraints predicted by neural networks and probability density functions. *Protein Engineering*, 10(11):1241–1248, 1997.

[140] Andreas R Gruber, Stephan H Bernhart, You Zhou, and Ivo L Hofacker. RNALfoldz: Efficient prediction of thermodynamically stable, local secondary structures. In *German Conference on Bioinformatics*, volume 173, pages 12–21, 2010.

[141] R. M. Fryer, J. Randall, T. Yoshida, L. L. Hsiao, J. Blumenstock, K. E. Jensen, T. Dimofte, R. V. Jensen, and S. R. Gullans. Global analysis of gene expression: methods, interpretation, and pitfalls. *Experimental Nephrology*, 10(2):64–74, 2002.

[142] Gajendra Raghava and Joon Han. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*, 6(1):59, 2005.

[143] Xian Wang, Ao Li, Zhaohui Jiang, and Huanqing Feng. Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, 7(1):32, 2006.

[144] E. Myasnikova, A. Samsonova, M. Samsonova, and J. Reinitz. Support vector regression applied to the determination of the developmental age of a Drosophila

embryo from its segmentation gene expression patterns. *Bioinformatics*, 18:S87–S95, 2002.

[145] Giorgio Guzzetta, Giuseppe Jurman, and Cesare Furlanello. A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinformatics*, 11(Suppl 8):S3, 2010.

[146] Qiang Liu, Kevin K. Lin, Bogi Andersen, Padhraic Smyth, and Alexander Ihler. Estimating replicate time shifts using gaussian process regression. *Bioinformatics*, 26(6):770–776, 2010.

[147] Wen Zhang, Ying wooi Wan, Genevera Allen, Kaifang Pang, Matthew Anderson, and Zhandong Liu. Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics*, 14:S7, 2013.

[148] J. Gui and H. Li. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

[149] Albert M. Levin, Debashis Ghosh, Kathleen R. Cho, and Sharon L. R. Kardia. A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, 21(12):2867–2874, 2005.

[150] Concha Bielza, Vctor Robles, and Pedro Larraaga. Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, 38(5):5110–5118, 2011.

[151] Pei-Chun Chen, Su-Yun Huang, Wei Chen, and Chuhsing Hsiao. A new regularized least squares support vector regression for gene selection. *BMC Bioinformatics*, 10(1):44, 2009.

[152] Dong Dong, Xiaojian Shao, Naiyang Deng, and Zhaolei Zhang. Gene expression variations are predictive for stochastic noise. *Nucleic Acids Research*, 2010.

[153] Roger Koenker. *Quantile regression.* Number 38. Cambridge University Press, 2005.

[154] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[155] Wouter G. Touw, Jumamurat R. Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha A. F. T. van Hijum. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, 14(3):315–326, 2013.

[156] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[157] V. N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[158] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[159] Sarunas Raudys. *Feature over-selection*, pages 622–631. Structural, Syntactic, and Statistical Pattern Recognition. Springer, 2006.

[160] Petr Somol, Ji Grim, Jana Novoviov, and Pavel Pudil. Improving feature selection process resistance to failures caused by curse-of-dimensionality effects. *Kybernetika*, 47(3):401–425, 2011.

[161] Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[162] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

[163] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[164] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[165] A. Torres and J. J. Nieto. Fuzzy logic in medicine and bioinformatics. *Journal of Biomedicine & Biotechnology*, 2006(2):91908, 2006.

[166] Maysam F. Abbod, Diedrich G. von Keyserlingk, Derek A. Linkens, and Mahdi Mahfouf. Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems*, 120(2):331–349, 2001.

[167] A Mesut Erzurumluoglu, Santiago Rodriguez, Hashem A Shihab, Denis Baird, Tom G Richardson, Ian NM Day, and Tom R Gaunt. Identifying highly penetrant disease causal mutations using next generation sequencing: Guide to whole process. *BioMed Research International*, 2015.

[168] Lotfi A Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[169] George Klir and Bo Yuan. *Fuzzy sets and fuzzy logic*, volume 4. Prentice Hall, 1995.

[170] Lotfi A Zadeh. The concept of a linguistic variable and its application to approximate reasoning-II. *Information Sciences*, 8(4):301–357, 1975.

[171] Ebrahim H. Mamdani and Sedrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-machine Studies*, 7(1): 1–13, 1975.

[172] Hans Hellendoorn and Christoph Thomas. Defuzzification in fuzzy controllers. *Journal of Intelligent and Fuzzy Systems*, 1(2):109–123, 1993.

[173] An introductory survey of fuzzy control. *Information Sciences*, 36(12):59 – 83, 1985.

[174] C.C. Lee. Fuzzy logic in control systems: fuzzy logic controller. II. *Systems, Man and Cybernetics, IEEE Transactions on*, 20(2):419–435, 1990.

[175] Eduardo Massad, Neli Regina Siqueira Ortega, Laecio Carvalho de Barros, and Claudio J. Struchiner. *Fuzzy logic in action: applications in epidemiology and beyond*, volume 232. Springer Science & Business Media, 2009.

[176] Dragan Kukolj. Design of adaptive Takagi-Sugeno-Kang fuzzy models. *Applied Soft Computing*, 2(2):89–103, 2002.

[177] Gorazd Karer and Igor Skrjanc, editors. *Predictive Approaches to Control of Complex Systems*, chapter Hybrid Fuzzy Model, pages 33–47. Springer, 2013.

[178] Chun-Tian Cheng, Jian-Yi Lin, Ying-Guang Sun, and Kwokwing Chau. Long-term prediction of discharges in manwan hydropower using adaptive-network-based fuzzy inference systems models. In *Advances in Natural Computation*, pages 1152–1161. Springer, 2005.

[179] Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-I. *Information Sciences*, 8(3):199–249, 1975.

[180] Hani A. Hagras. A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots. *Fuzzy Systems, IEEE Transactions on*, 12(4):524–539, 2004.

[181] Christian Wagner and Hani Hagras. A genetic algorithm based architecture for evolving type-2 fuzzy logic controllers for real world autonomous mobile robots. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–6, 2007.

[182] John Figueroa, Jorge Posada, Jairo Soriano, Miguel Melgarejo, and Sergio Rojas. A type-2 fuzzy controller for tracking mobile objects in the context of robotic soccer games. In *Fuzzy Systems (FUZZ'05), The 14th IEEE International Conference on*, pages 359–364, 2005.

[183] Simon Coupland, Mario Augusto Gongora, Robert John, and K. Wills. A comparative study of fuzzy logic controllers for autonomous robots. In *Proceedings of the Information Processing and Management of Uncertainty in Knowledge-based Systems Conference*, pages 1332–1339, 2006.

[184] Roberto Sepulveda, Oscar Castillo, Patricia Melin, Antonio Rodriguez-Diaz, and Oscar Montiel. Experimental study of intelligent controllers under uncertainty using type-1 and type-2 fuzzy logic. *Information Sciences*, 177(10):2023–2048, 2007.

[185] Dongrui Wu and Woei Wan Tan. A type-2 fuzzy logic controller for the liquid-level process. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, volume 2, pages 953–958, 2004.

[186] J. M. Mendel and R. I. John. Type-2 fuzzy sets made simple. *Fuzzy Systems, IEEE Transactions on*, 10(2):117–127, 2002.

[187] Jerry M. Mendel, R. I. John, and Feilong Liu. Interval type-2 fuzzy logic systems made simple. *Fuzzy Systems, IEEE Transactions on*, 14(6):808–821, 2006.

[188] R. I. John, Jerry Mendel, and Jenny Carter. The extended sup-star composition for type-2 fuzzy sets made simple. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1441–1445, 2006.

[189] Jerry M Mendel. On the importance of interval sets in type-2 fuzzy logic systems. In *Proceedings of the 9th Joint IFSA World Congress and 20th NAFIPS International Conference*, volume 3, pages 1647–1652, 2001.

[190] Qilian Liang and Jerry M. Mendel. Interval type-2 fuzzy logic systems: theory and design. *Fuzzy Systems, IEEE Transactions on*, 8(5):535–550, 2000.

[191] J. M. Mendel. *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Prentice Hall, 2001.

[192] Nilesh N Karnik, Jerry M Mendel, and Qilian Liang. Type-2 fuzzy logic systems. *Fuzzy Systems, IEEE Transactions on*, 7(6):643–658, 1999.

[193] Dongrui Wu and Jerry M Mendel. Enhanced Karnik-Mendel algorithms. *Fuzzy Systems, IEEE Transactions on*, 17(4):923–934, 2009.

[194] Dongrui Wu and Jerry M Mendel. Enhanced Karnik-Mendel algorithms for Interval Type-2 fuzzy sets and systems. In *Fuzzy Information Processing Society (NAFIPS'07), Annual Meeting of the North American*, pages 184–189. IEEE, 2007.

[195] Thomas A. Runkler and James C. Bezdek. Function approximation with polynomial membership functions and alternating cluster estimation. *Fuzzy Sets and Systems*, 101(2):207–218, 1999.

[196] T. A. Runkler and J. C. Bezdek. Alternating cluster estimation: a new tool for clustering and function approximation. *Fuzzy Systems, IEEE Transactions on*, 7 (4):377–393, 1999.

[197] Hui Cao, Lixin Jia, Gangquan Si, and Yanbin Zhang. A clustering-analysis-based membership functions formation method for fuzzy controller of ball mill pulverizing system. *Journal of Process Control*, 23(1):34–43, 2013.

[198] Raghu Krishnapuram. Generation of membership functions via possibilistic clustering. In *Proceedings of the Third IEEE Conference on Fuzzy Systems*, pages 902–908, 1994.

[199] JR Boston. Effects of the shape of fuzzy membership functions on fuzzy inference. In *Proceedings of the Third International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society (ISUMA-NAFIPS'95)*, pages 32–37, 1995.

[200] Sanya Mitaim and Bart Kosko. The shape of fuzzy sets in adaptive function approximation. *Fuzzy Systems, IEEE Transactions on*, 9(4):637–656, 2001.

[201] Liang Wang and R. Langari. Complex systems modeling via fuzzy logic. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 26(1):100–106, 1996.

[202] R. R. Yager and D. P. Filev. Unified structure and parameter identification of fuzzy models. *Systems, Man and Cybernetics, IEEE Transactions on*, 23(4):1198–1205, 1993.

[203] Dongrui Wu. Twelve considerations in choosing between Gaussian and trapezoidal membership functions in interval type-2 fuzzy logic controllers. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–8, 2012.

[204] Dongrui Wu and Woei Wan Tan. Genetic learning and performance evaluation of interval type-2 fuzzy logic controllers. *Engineering Applications of Artificial Intelligence*, 19(8):829 – 841, 2006.

[205] Woei Wan Tan, Chek Liang Foo, and Teck Wee Chua. Type-2 fuzzy system for ECG arrhythmic classification. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–6, 2007.

[206] Robert Fuller. *Introduction to neuro-fuzzy systems*, volume 2. Springer, 2000.

[207] Jyh-Shing Roger Jang, Chuen-Tsai Sun, and Eiji Mizutani. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review]. *Automatic Control, IEEE Transactions on*, 42(10):1482–1484, 1997.

[208] Detlef Nauck, Frank Klawonn, and Rudolf Kruse. *Foundations of neuro-fuzzy systems*. John Wiley & Sons, 1997.

[209] J. S R. Jang. ANFIS: adaptive-network-based fuzzy inference system. *Systems, Man and Cybernetics, IEEE Transactions on*, 23(3):665–685, 1993.

[210] David E. Goldberg. Genetic algorithms in search, optimization, and machine learning. *Addion-Wesley*, 1989.

[211] Plamen P. Angelov. *Evolving rule-based models: a tool for design of flexible adaptive systems*, volume 92. Springer, 2002.

[212] Oscar Cordon. *Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases*, volume 19. World Scientific, 2001.

[213] Francisco Herrera and Jose Luis Verdegay. *Genetic algorithms and soft computing.* Physica-Verlag, 1996.

[214] Witold Pedrycz. *Fuzzy evolutionary computation.* Kluwer Academic Publishers, 1997.

[215] Elie Sanchez, Takanori Shibata, and Lotfi Asker Zadeh. *Genetic algorithms and fuzzy logic systems: Soft computing perspectives*, volume 7. World Scientific, 1997.

[216] Majid Almaraashi, Robert John, Simon Coupland, and Adrian Hopgood. Time series forecasting using a TSK fuzzy system tuned with simulated annealing. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–6, 2010.

[217] Majid Almaraashi and Robert John. Tuning of type-2 fuzzy systems by simulated annealing to predict time series. In *Proceedings of the World Congress on Engineering*, volume 2, 2011.

[218] A. Chervonenkis V. N. Vapnik. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

[219] V. N. Vapnik. *The nature of statistical learning theory.* Springer, 1995.

[220] V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.

[221] Harris Drucker, C. J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir N. Vapnik. *Support Vector Regression Machines*, volume 9 of *Advances in Neural Information Processing Systems.* MIT Press, 1996.

[222] K-R Müller, Alex J Smola, Gunnar Rätsch, Bernhard Schölkopf, Jens Kohlmorgen, and Vladimir Vapnik. Predicting time series with support vector machines. In *Artificial Neural Networks (ICANN'97)*, pages 999–1004. Springer, 1997.

[223] Mark O. Stitson, Alex Gammerman, Vladimir Vapnik, Volodya Vovk, Chris Watkins, and Jason Weston. Advances in kernel methods. chapter Support Vector Regression with ANOVA Decomposition Kernels, pages 285–291. MIT Press, 1999.

[224] Davide Mattera and Simon Haykin. Advances in kernel methods. chapter Support Vector Machines for Dynamic Reconstruction of a Chaotic System, pages 211–241. MIT Press, 1999.

[225] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[226] Alex J. Smola and Bernhard Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[227] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[228] Udo Seiffert, Barbara Hammer, Samuel Kaski, and Thomas Villmann. Neural networks and machine learning in bioinformatics-theory and applications. In *European Symposium on Artificial Neural Networks (ESANN)*, pages 521–532, 2006.

[229] BS Everitt, S Landau, and M Leese. *Cluster Analysis*. Arnold, 2001.

[230] Rajesh N. Dave and Raghuram Krishnapuram. Robust clustering methods: a unified view. *Fuzzy Systems, IEEE Transactions on*, 5(2):270–293, 1997.

[231] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

[232] Rui Xu and D. Wunsch II. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[233] Hichem Frigui and Raghu Krishnapuram. Clustering by competitive agglomeration. *Pattern Recognition*, 30(7):1109–1119, 1997.

[234] Raghu Krisnapuram James C. Bezdek, James Keller and Nikhil R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4. Springer, 2005.

[235] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[236] Richard O Duda and Peter E Hart. *Pattern Recognition and Scene Analysis*. Wiley, 1973.

[237] Timothy C. Havens, J. C. Bezdek, James M. Keller, and Mihail Popescu. Clustering in ordered dissimilarity data. *International Journal of Intelligent Systems*, 24(5):504–528, 2009.

[238] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.

[239] Stuart Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[240] Enrique H. Ruspini. A new approach to clustering. *Information and control*, 15 (1):22–32, 1969.

[241] James C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[242] James C Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(2):191–203, 1984.

[243] S Nascimento, B Mirkin, and F Moura-Pires. A fuzzy clustering model of data and fuzzy c-means. In *Fuzzy Systems (FUZZ-IEEE), The Ninth IEEE International Conference on*, volume 1, pages 302–307, 2000.

[244] Richard J Hathaway, GW Rogers, and JC Bezdek. Random vector clustering using fuzzy c-means. In *Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, pages 251–255, 1998.

[245] Mohamed Fadhel Saad and Adel M. Alimi. Modified fuzzy possibilistic C-means. In *Proceedings of the international multi conference of engineers and computer scientists*, volume 1, pages 18–20, 2009.

[246] Miin-Shen Yang and Kuo-Lung Wu. A possibilistic type of alternative fuzzy C-means. In *Fuzzy Systems (FUZZ-IEEE'02), Proceedings of the IEEE International Conference on*, volume 2, pages 1456–1459, 2002.

[247] I Sledge, J Bezdek, T Havens, and J Keller. A relational dual of the fuzzy possibilistic c-means algorithm. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–9, 2010.

[248] Nikhil R Pal, Kuhu Pal, James M Keller, and James C Bezdek. A possibilistic fuzzy c-means clustering algorithm. *Fuzzy Systems, IEEE Transactions on*, 13(4): 517–530, 2005.

[249] Radha P. Sandhir and Satish Kumar. Dynamic fuzzy c-means (dFCM) clustering for continuously varying data environments. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1–8, 2010.

[250] Marc Roubens. Pattern classification problems and fuzzy sets. *Fuzzy sets and systems*, 1(4):239–253, 1978.

[251] Richard J Hathaway, John W Davenport, and James C Bezdek. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, 22(2):205–212, 1989.

[252] David Wishart. 256. Note: An algorithm for hierarchical classifications. *Biometrics*, pages 165–170, 1969.

[253] Godfrey N. Lance and William Thomas Williams. A general theory of classificatory sorting strategies I. Hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967.

[254] Robin Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[255] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[256] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[257] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[258] Michael Steinbach, Levent Ertz, and Vipin Kumar. *The challenges of clustering high dimensional data*, pages 273–309. New Directions in Statistical Physics. Springer, 2004.

[259] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. *When is "nearest neighbor" meaningful?*, pages 217–235. Database Theory. Springer, 1999.

[260] Nikhil R Pal and James C Bezdek. On cluster validity for the fuzzy c-means model. *Fuzzy Systems, IEEE Transactions on*, 3(3):370–379, 1995.

[261] Daniele Soria and Jonathan M Garibaldi. A novel framework to elucidate core classes in a dataset. In *Evolutionary Computation (CEC), IEEE Congress on*, pages 1–8, 2010.

[262] Malay K Pakhira and Amrita Dutta. Determination of number of clusters using VAT images and genetic algorithms. In *Emerging Applications of Information Technology (EAIT), Second International Conference on*, pages 357–360, 2011.

[263] Malay K Pakhira and Amrita Dutta. Finding number of clusters using VAT image, PBM index and genetic algorithms. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pages 217–221, 2010.

[264] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. IEEE International Conference on Neural Networks*, volume IV, pages 1942–1948, 1995.

[265] Joseph C Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 1973.

[266] James C Bezdek and Richard J Hathaway. VAT: A tool for visual assessment of (cluster) tendency. In *Neural Networks (IJCNN'02), Proceedings of the International Joint Conference on*, volume 3, pages 2225–2230, 2002.

[267] Timothy C. Havens and James C. Bezdek. An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm. *Knowledge and Data Engineering, IEEE Transactions on*, 24(5):813–822, 2012.

[268] Zheng Zhao, Fred Morstatter, Shashvata Sharma, Salem Alelyani, Aneeth Anand, and Huan Liu. Advancing feature selection research. *ASU Feature Selection Repository*, 2010.

[269] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, 2005.

[270] Petr Somol, Jana Novovicová, and Pavel Pudil. Efficient feature subset selection and subset size optimization. *Pattern Recognit Recent Adv*, 2010.

[271] Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Springer, 1998.

[272] Pabitra Mitra, CA Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002.

[273] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 333–342, 2010.

[274] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

[275] Lior Wolf and Amnon Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *The Journal of Machine Learning Research*, 6:1855–1887, 2005.

[276] Mohd Nayeem, Ashique Ridwan, Md A Mannan Joadder, Shahrin Ahammad Shetu, Farzin Raeeda Jamil, and Abdullah Al Helal. Feature selection for breast cancer detection from ultrasound images. In *Informatics, Electronics & Vision (ICIEV), International Conference on*, pages 1–6, 2014.

[277] Martin Stahl, Wolfgang Guba, and Manfred Kansy. Integrating molecular design resources within modern drug discovery research: the roche experience. *Drug Discovery Today*, 11(7):326–333, 2006.

[278] Karsten Wicke and Javier Garcia-Ladona. The dopamine D3 receptor partial agonist, BP 897, is an antagonist at human dopamine D3 receptors and at rat somatodendritic dopamine D3 receptors. *European Journal of Pharmacology*, 424 (2):85–90, 2001.

[279] SK Singh, N Dessalew, and PV Bharatam. 3D-QSAR CoMFA study on indenopyrazole derivatives as cyclin dependent kinase 4 (CDK4) and cyclin dependent kinase 2 (CDK2) inhibitors. *European Journal of Medicinal Chemistry*, 41(11): 1310–1319, 2006.

[280] Chandrabose Selvaraj, Sunil Kumar Tripathi, Karnati Konda Reddy, and Sanjeev Kumar Singh. Tool development for prediction of pIC50 values from the IC50 values - A pIC50 value calculator. *Current Trends in Biotechnology and Pharmacy*, 5(2):1104–1109, 2011.

[281] Nigus Dessalew and Sanjeev K Singh. 3D-QSAR CoMFA and CoMSIA study on benzodipyrazoles as cyclin dependent kinase 2 inhibitors. *Medicinal Chemistry*, 4 (4):313–321, 2008.

[282] Xiao-Yun Lu, Ya-Dong Chen, Ni-yue Sun, Yong-Jun Jiang, and Qi-Dong You. Molecular-docking-guided 3D-QSAR studies of substituted isoquinoline-1, 3-(2H, 4H)-diones as cyclin-dependent kinase 4 (CDK4) inhibitors. *Journal of Molecular Modeling*, 16(2):163–173, 2010. Springer.

[283] R. G. D. Steel, J. H. Torrie, and D. A. Dickey. *Principles and Procedures of Statistics*. McGraw-Hill, 1997.

[284] Jerome L. Myers and Arnold D. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, 2nd edition, 2003.

[285] Ovidiu Ivanciuc. Comparative evaluation of prediction algorithms (CoEPrA), 2006. URL `http://www.coepra.org/`.

[286] Channa K Hattotuwagama, Pingping Guan, Irini A Doytchinova, and Darren R Flower. New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity. *Organic & Biomolecular Chemistry*, 2(22):3274–3283, 2004.

[287] Soichi Tanabe. Epitope peptides and immunotherapy. *Current Protein and Peptide Science*, 8(1):109–118, 2007.

[288] A Bossi, F Bonini, APF Turner, and SA Piletsky. Molecularly imprinted polymers for the recognition of proteins: the state of the art. *Biosensors and Bioelectronics*, 22(6):1131–1137, 2007.

[289] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(suppl 1):D202–D205, 2008.

[290] I. A. Doytchinova, M. J. Blythe, and D. R. Flower. Additive method for the prediction of protein-peptide binding affinity. application to the MHC class I molecule HLA-A*0201. *J Proteome Res*, 1(3):263–272, 2002.

[291] Charles Bergeron, Theresa Hepburn, C Matthew Sundling, Michael Krein, Bill Katt, Nagamani Sukumar, Curt M Breneman, and Kristin P Bennett. Prediction of peptide bonding affinity: kernel methods for nonlinear modeling. *arXiv:1108.5397*, 2011.

[292] Atulji Srivastava, Shameek Ghosh, N Anantharaman, and VK Jayaraman. Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests. *Journal of Immunological Methods*, 387(1):284–292, 2013.

[293] Yi-Cheng Chen, N. R. Pal, and I-Fang Chung. An integrated mechanism for feature selection and fuzzy rule extraction for classification. *Fuzzy Systems, IEEE Transactions on*, 20(4):683–698, 2012.

[294] Jung-Hsien Chiang and Pei-Yi Hao. Support vector learning mechanism for fuzzy rule-based modeling: a new approach. *Fuzzy Systems, IEEE Transactions on*, 12 (1):1–12, 2004.

[295] Julio César Tovar. Fuzzy neural modeling via clustering and support vector machines. In *Control Applications (CCA), IEEE International Conference on*, pages 24–29, 2007.

[296] Yixin Chen and J. Z. Wang. Support vector learning for fuzzy rule-based classification systems. *Fuzzy Systems, IEEE Transactions on*, 11(6):716–728, 2003.

[297] J. M. Leski. TSK-fuzzy modeling based on e-insensitive learning. *Fuzzy Systems, IEEE Transactions on*, 13(2):181–193, 2005.

[298] Chia-Feng Juang, Shih-Hsuan Chiu, and Shen-Jie Shiu. Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(6):1077–1087, 2007.

[299] Chia-Feng Juang, Cheng-Da Hsieh, and Jyun-Lang Hong. Fuzzy clustering-based neural fuzzy network with support vector regression. In *Industrial Electronics and Applications (ICIEA), The 5th IEEE Conference on*, pages 576–581, 2010.

[300] Chia-Feng Juang and Cheng-Da Hsieh. A fuzzy system constructed by rule generation and iterative linear SVR for antecedent and consequent parameter optimization. *Fuzzy Systems, IEEE Transactions on*, 20(2):372–384, 2012.

[301] J.-S R. Jang and Chuen-Tsai Sun. Neuro-fuzzy modeling and control. *Proceedings of the IEEE*, 83(3):378–406, 1995.

[302] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[303] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985.

[304] RR Matheson and Harold A. Scheraga. Calculation of the Zimm-Bragg cooperativity parameter from a simple model of the nucleation process. *Macromolecules*, 16(7):1037–1043, 1983.

[305] Yudong Cai, Tao Huang, Lele Hu, Xiaohe Shi, Lu Xie, and Yixue Li. Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids*, 42(4):1387–1395, 2012.

[306] Dingfang Li and Wenchao Hu. A relevance vector machine based quantitative prediction method for mouse class I MHC peptide binding affinity. In *International Conference on Biomedical and Pharmaceutical Engineering. ICBPE*, pages 349–353, 2006.

[307] I. G. Sharina, R. Zhao, Y. Wang, S. Babani, and I. D. Goldman. Mutational analysis of the functional role of conserved arginine and lysine residues in transmembrane domains of the murine reduced folate carrier. *Molecular Pharmacology*, 59(5):1022–1028, 2001.

[308] R. I. John. Type 2 fuzzy sets: an appraisal of theory and applications. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(06): 563–576, 1998.

[309] Nilesh Naval Karnik and Jerry M Mendel. Type-2 fuzzy logic systems: type-reduction. In *Systems, Man, and Cybernetics, IEEE International Conference on*, volume 2, pages 2046–2051, 1998.

[310] N. N. Karnik and Jerry M. Mendel. Centroid of a type-2 fuzzy set. *Information Sciences*, 132(1):195–220, 2001.

[311] Xinwang Liu, Jerry M Mendel, and Dongrui Wu. Study on enhanced Karnik-Mendel algorithms: Initialization explanations and computation improvements. *Information Sciences*, 184(1):75–91, 2012.

[312] Chia-Feng Juang, Ren-Bo Huang, and Wei-Yuan Cheng. An interval type-2 fuzzy-neural network with support-vector regression for noisy regression problems. *Fuzzy Systems, IEEE Transactions on*, 18(4):686–699, 2010.

[313] Qilian Liang and J. M. Mendel. Equalization of nonlinear time-varying channels using type-2 fuzzy adaptive filters. *Fuzzy Systems, IEEE Transactions on*, 8(5): 551–563, 2000.

[314] Hongwei Wu and Jerry M Mendel. Introduction to uncertainty bounds and their use in the design of interval type-2 fuzzy logic systems. In *Fuzzy Systems (FUZZ-IEEE), The 10th IEEE International Conference on*, volume 2, pages 662–665, 2001.

[315] S. Coupland and R. John. Geometric type-1 and type-2 fuzzy logic systems. *Fuzzy Systems, IEEE Transactions on*, 15(1):3–15, 2007.

[316] Sarah Greenfield, Francisco Chiclana, Simon Coupland, and Robert John. The collapsing method of defuzzification for discretised interval type-2 fuzzy sets. *Information Sciences*, 179(13):2055–2069, 2009.

[317] Maowen Nie and Woei Wan Tan. Towards an efficient type-reduction method for interval type-2 fuzzy logic systems. In *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*, pages 1425–1432, 2008.

[318] M. Biglarbegian, W. W. Melek, and J. M. Mendel. On the stability of interval type-2 TSK fuzzy logic control systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(3):798–818, 2010.

[319] Michio Sugeno and Takahiro Yasukawa. A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 1(1):7–31, 1993.

[320] Oscar Cordón, Francisco Herrera, and Pedro Villar. Analysis and guidelines to obtain a good uniform fuzzy partition granularity for fuzzy rule-based systems using simulated annealing. *International Journal of Approximate Reasoning*, 25 (3):187–215, 2000.

[321] María José Gacto, Rafael Alcalá, and Francisco Herrera. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360, 2011.